様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月１１日現在

研究種目：基盤研究（C）
研究期間：2007～2009
課題番号：19500076
研究課題名（和文）　圧縮空間を用いたマルチメディアデータマイニングとそのウェブマイニングへの応用
研究課題名（英文）　Compression feature space based data mining and its application to web mining

研究代表者
渡邊　俊典　（WATANABE TOSHINORI）
電気通信大学・大学院情報システム学研究科・教授
研究者番号：10242348

研究成果の概要（和文）：
インターネットや携帯電話の発展の中で文章、音声、画像などのマルチメディアデータが爆発的に増大している。本研究では、人手介入なしに、計算機によってこれらを分類あるいは検索する方式を検討した。我々が過去検討してきた、圧縮率によるテキストの特徴表現方式を原理としつつ、より高性能な圧縮性特徴空間の構成可能性の検討、文書や画像分類への適用などを試みた。文法知識を事前準備せずに、文章や画像に適用できること、従来方式をしのぐ性能も発揮できることなどを確認できた。なお、EU での衛星画像利用地球環境管理国際プロジェクト（GEOSS）関係機関より招待され、衛星画像処理への応用可能性について講演も実施した。

研究成果の概要（英文）：
Rapid explosion of internet and mobile telephone require us highly automatic computerized systems that can classify and/or recognize multimedia data including documents, sounds, and images. In this project, based on the principle of text featuring by compressibility which we have been exploiting, new investigations are made on the new problem of how to construct efficient compressibility feature space and its applications. Compared with conventional approaches, its merits include: free of linguistic knowledge (grammar), wide applicability, simplicity, and high performance.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	800,000	240,000	1,040,000
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,400,000	420,000	1,820,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学、メディア情報学･データベース
キーワード：マルチメディア情報処理、データマイニング、データ圧縮

１．研究開始当初の背景

近年、情報化が急速に発展するにつれ、計算機の扱うデータの大容量化、種類の多様化が進んでいる。データベースに保存されたデータと比べ、インターネットや携帯電話で利用されるテキスト、画像、音声などは構文木

などの構造化手法になじみにくく、これらを、文法規則によってモデル化しようとする検討もなされているが、期待されたほどには成功していない。これらの構造化しにくいデータへの柔軟なアクセスを実現することは未解決の課題である。現状では主要な取り組みは、下記のようである。

マルチメディアデータの、①意味ラベルを人が付与し、以降の計算機処理を容易化するアプローチと、②なるべく人手によらず計算機で構造や意味を自動把握することを目指すデータマイニングアプローチとに大別される。前者は、多様な意味を持つマルチメディアデータに特定のラベルを与えることが将来の検索を限定してしまうという問題や、情報付与のための人手コストの問題を含んでいるが、これといった代替手段がないため広く採用されている。後者は、近年多くの研究がなされているにもかかわらず、顕著な応用領域を切り開くまでにはいたっていない。例えば、ベイジアンネットワークや Markov ランダム場（MRF）でデータの内容をモデル化する手法が近年盛んであるが、モデルの基本構造は人による設計を要し、またモデルによるデータの解析には高い計算コストが必要となる。

２．研究の目的

本研究では、圧縮率特徴空間（圧縮空間、あるいは PRDC（Pattern Representation Scheme Using Data Compression)）における多次元データ表現方式を利用した新たなマルチメディアデータマイニングの可能性を探求する。その可能性を左右すると予想される下記の諸機能の検討を目的とする。

（1）テキスト解析の検討:PRDC 研究の中で、課題となっているテキストの内容分析に関し、文法を用いず、類似性解析・検索、断片の特性分析（不要語、新語などの摘出)、要約を可能とする方式。および、時間変化するテキスト集合への本手法の対応能力を向上させるための自動適応方式、分析結果の人への視覚化表示方式を検討する。

（２）画像解析の検討：画像に対する、類似性解析・検索、特徴的領域の同定、要約方式を検討する。

（３）マルチメディアデータ解析：圧縮性特徴量とデータの共起性を利用した、異タイプメディアデータ間の自動関連付けの可能性を検討する。

（４）Web マイニングシステムのプロトタイピング：上記の諸手法を統合的に用いて、Web 情報の自動分類、検索、内容推測、結果のビジュアル表示などを行なうプロトタイプシステムを Web サーバ上に実現し、現在の先端検索エンジンと比較し、問題点を発見して本研究を更に改良する。

３．研究の方法

（１）テキスト解析の検討：テキスト全体の圧縮性による類似度解析とともに、テキスト断片の圧縮性に基づくトピック抽出を試みた。提案手法を検証するために、人工のモデル文書を生成し、原理レベルでの検証を行なった後、実テキストによる検証を行なった。流行語など、テキストの時間変化については新トピックの出現問題として扱った。また、解析の基礎となる圧縮空間の構成方式に関し、分類能力を保持したままでの次元低下法などを検討した。

（２）画像解析の検討：画像をランレングス方向（行方向）に走査してコディングして得られるテキストを用い、上記（１）を適用することを試みた。さらに、テクスチャ画像について、向きの変化やズームに対するロバスト性も併せて検討した。

（３）マルチメディアデータ解析：上記（２）の画像解析研究の結果を踏まえ、Web データによく見られるような、画像とテキストの隣接配置に対する本方式の適用法を考察することとした。

（４）Web マイニングシステムプロトタイピング：Web クローラで収集した実文書（日本語、英語、中国語、およびこれらの混在）について、上記（１）の応用可能性を検証し、既存手法と比較した。

（５）その他：上記研究の中で、新たに発掘される問題については、重要度に応じた検討を実施することとした。

４．研究成果

以下に、研究目的にリンクさせつつ、研究成果を説明する。

（１）テキスト解析の検討については、主として独立成分分析を利用した圧縮空間の独立化、文書間の関係分析方式、対話形式による多次元空間の次元削減方式について下記の成果を得た。

まず、独立圧縮空間の構成方式について検討した。特徴表現方法として有名な bag-of-words 法と N-gram モデルが、自然言語処理、テキストマイニングと文書解析によく利用されてきた。一方、PRDC はテキストデータのみではなく、マルチメディアデータ解析において、良い性能を示したが、圧縮辞書の選択や特徴空間の構築などの問題が残されている。本研究では、独立な圧縮率空間の構築法を検討した。

提案法では、独立成分分析法を利用して、圧縮率空間の次元数を削減し、より性能がよい特徴表現空間を構築に成功した。実験では、日本語のニュースサイトから収集したデータと、URCS、CLASSIC3、Reuters-21578 とい

うテキスト解析によく利用されているベンチマークを用いて、従来法の bag-of-words 法と N-gram モデルとの性能比較を行った。実験の結果、提案法を利用したテキストクラスタリング結果は従来の手法より良好であった（学会発表②と③）。

また、文書間の関係を分析する研究は、自然言語処理に基づいたものが主流となっている。単語出現頻度を利用する場合、文書中の単語の出現頻度をカウントする。これだけでも文書の特徴を把握することが可能となる。更に、出現頻度をベクトルとして表し、ベクトル間の類似度を算出することで、文書間の類似性を調べることができる。キーワード抽出を利用する場合、文書中の代表的なキーワードを抽出することにより、類似文書の判定はもちろん、ウェブページ検索や要約文抽出などに応用することができる。代表的な手法に TF-IDF がある。この処理の基礎として、形態素解析が行われている。

日本語の形態素解析は、実用レベルに達するまで進んでいるが、複合語、未知語の扱いや、品詞情報の利用や、ストップワードの扱いや辞書の品質や、安定かつ高速なエンジンの開発などの問題点も残されている。特に、常に新しい単語が生まれるインターネットなどの情報処理としては、辞書に登録されていない全くの未知の用語に対して適応性が欠けていることや、複合語の抽出単位に曖昧性があることなどが挙げられる。

本研究では、自然言語処理を用いず、文書間の関係を分析、提示する簡素かつ汎用性のある手法を検討した。

具体的には文書間のトピックの重なりの関係に注目し、データの圧縮率を用いることで、文書間の関係を分析する方法を検討した。まず、意図的に文書間の関係を構成できるモデル文書を導入し、これを提案手法の実験、検証に用いた。実際の文書中のトピックは、読み手の解釈やその目的によって多種多様であり、トピックの完全に客観的な抽出は困難である。モデル文書を生成することで、提案手法の関係分析能力の検証を効率化することができた（図 1）。この後、実際の文書に対し実験を行い、提案手法の有効性を確認した。圧縮率に注目することで、類似文書のクラスタリング、不要語の除去、共通トピックの把握など、すべて同じフレームワークで実現する可能性を原理レベルではあるが、確認できた（図２）。（学会発表⑥と⑧）

	逢葵渥旭悪 monkey 逢愛 monkey 阿nkey 始始葵娃 cankey 渥茜茜茜亜梓哀芦幕芦茜葵 monk葵葵梓 monkey 娃芦 monkey 葵逢幕梓鯵芦娃娃 monkeg 幕哀梓幕茜葵悪握 monkey 葵唖梓阿逢握 monkey 挨 horse 娃芦梓悪悪旭娃娃旭始亜梓娃梓 monkey 愛幕唖茜握挨幕哀鯵芦茜 monkey	娃茜茜唖 eagle 娃旭阿娃悪俺俺幕 dog eagle 愛 monk茜 eagle 旭亜幕娃亜葵始 eagle 亜鯵渥哀幕哀逢阿渥 eag茜 eagle 俺娃葵鯵eagle 渥葵旭唖旭始愛 eagle 握唖挨agle 唖渥握梓哀葵鯵逢茜 eagle 芦渥亜阿 eagle 旭俺俺悪鯵逢唖唖渥愛 owl挨挨逢葵逢逢唖 eag哀始阿愛 eagle 阿	t 阿亜 cat 唖幕愛唖幕阿芦握旭 cat 逢鯵梓葵哀茜 cat 俺愛握握葵悪幕俺握 cat幕旭愛芦悪 cat 梓梓俺愛始愛 cat 哀茜愛cat 握始旭逢幕 do挨渥 cat 悪唖鯵愛哀悪旭逢 cat 阿哀挨 h梓 cat 梓渥阿握俺俺nkey 渥娃始握梓渥葵逢逢幕握梓阿 cat
A	O	O	O
B	O	O	X
C	O	X	O

図 1. モデル文書を利用した文書間関係分析結果

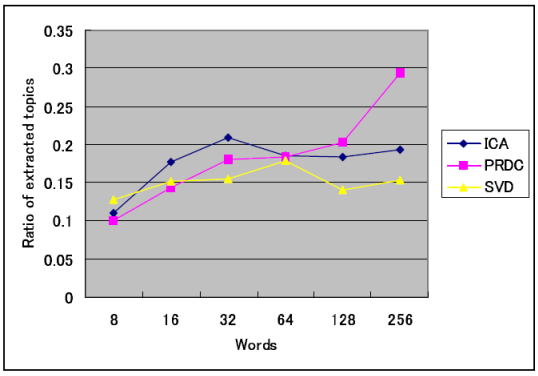

図 2. 大規模データにおけるトピック抽出率の比較（提案法、ICA、SVD）

次に、高次元特徴空間を用いるテキスト分類等において不必要な次元軸を排除することは計算量などの面から重要な問題であると認識し、この問題を解決するためには不要と考えられる次元を見出し、分類性能を保つ範囲でその次元を削除していく方式を検討した。まずテキストをその圧縮率ベクトルで特徴付け、次元縮小法と次元縮小に必要なパラメータ設定を支援する指標関数を用いて対話的に次元縮小を行う方式を得た。分類精度を保ったまま約 50%の次元縮小が可能となった。

表１．K-means 法の分類結果

人手による分類		k-means 法により得られたクラス							
トピック	記事数	クラス1	クラス2	クラス3	クラス4	クラス5	クラス6	クラス7	クラス8
履修	10	3					6	1	
松坂	10		10						
日興	10			4				6	
中越	10			1	9				
アジア杯	10					10			
参院選	20			1			16	3	
石川	10					10			
年金	20			1				10	9
分類再現率		30	100	40	90	*	80	*	45
分類精度		100	100	59	100	*	73	*	100

提案手法が、クラスタ数をあらかじめ設定することなく自動分類を行えること、更に分類能力を低化させずに次元縮小を行えることなどを示せた。分類性能(表 1、2)で示したように、k-means 法の分類再現率、精度よりも全てのクラスで提案手法が優れており、表 1 の「＊」のクラスのようにどのトピックのクラスであるか判定不可能なものが提案手法ではほとんどなく、トピックに沿った分類ができた。（学会発表⑨）

表２．次元縮小後の分類結果

人手による分類		次元縮小後の分類クラス							
トピック	記事数	クラス1	クラス2	クラス3	クラス4	クラス5	クラス6	クラス7	クラス8
履修	10	10							
松坂	10		10						
日興	10			10					
中越	10				7				
アジア杯	10					9			
参院選	20						15		
石川	10							9	
年金	20						2		17
分類再現率		100	100	100	70	90	75	90	85
分類精度		100	100	100	100	100	88	100	100

高次特徴空間の構成について、更に以下の検討も行った。すなわち、様々な文書を分類するための高次元特徴空間を定義したとき、その特徴空間の良さを計測したい。例えば物理学書は数学的な記述を多く含むので、そのクラスタ(集まり)は特徴空間上で歴史書や芸術書のクラスタよりは数学書のクラスタの近くに配されるはずであるが、実際にそうなっているか。他にもクラスタ間の距離が大きすぎたり小さすぎたりしないかなどを計測したい。こういった目的を達する最も直感的な方法は視覚的に特徴空間を観察することである。特徴空間は一般に非常に高次元である一方で、4 次元以上の空間を平面的な紙の上やモニターに描くことは困難である。このような場合、従来では高次元特徴空間を主成分分析や FastMap などの次元縮約法を用いて 2 次元、あるいは 3 次元空間に変換することが試みられている。

本研究では高次元特徴空間の形状を測り、無用な次元を削減するために、高次元空間内の多数の点の散布状況を調べるための外郭点の高速計算アルゴリズム（球分割法）を検討した。考案した球分割法は単純ガウス分布だけではなく様々な混合ガウス分布から外郭を取り出すことができることを確認した。また、単純ガウス分布に対する球分割法の実行時間は確率アルゴリズム（LSH）を用いることで 10 倍から 20 倍程度高速化できた。しかし、高速化の度合いは理論値よりも小さかった。この原因はデータの分布がアルゴリズムで仮定した球対称ではないためと考えられる。(学会発表⑤)

（２）画像解析の検討については、ランレングスコーディングによって得たテキストを用いることで原画像の分類を試みた。また、テクスチャ画像にも適用し、ある程度のサイズや角度変化の下でも良好な分類性能を得ることを確認した。

最初に、低周波成分主体の画像、高周波成分主体の画像を意図的に混合した人工画像や自然画像を用いて、画像をテキスト化する際に必要となるアルファベット定義用の単位ランレングス長 L の適正値を設定した。100 種類 1,000 枚の画像データセットに対してテキスト化画像の分類を試みたところ、平均 F 値尺度が 37%であり高精度とは言えない結果となった。原因として主に画像内の低周波背景が問題であると考えられた。画像が周波数成分の異なる領域を含む場合、単一の単位ランレングス長 L を用いると、情報損失が発生する。これに対しては、同一画像を複数の L でテキスト化し、得られる圧縮率ベクトルを結合することで、分類能力が若干高められることを確かめた（表３）。背景を除き、前景を取り出した場合についても同様の実験を行った。この場合、分類精度はより大きく向上した（表４）(学会発表①と⑦)。

表３．L=2,8(画素)の結合空間を用いた 10 種類データの分類結果

人手による分類		k-means 法により得られたクラス									
データセット		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
種類	画像数	バラ	ヒマワリ	象	日の出	地球	パンダ	トラ	岩	ヨット	家
バラ	10	**9**		1							
ヒマワリ	10	1	**5**	1			1	2			
象	10	3	3	**3**							1
日の出	10	8			**2**						
地球	10	1	1	3		**5**					
パンダ	10					1	**6**		1		2
トラ	10					1	3	**6**			
岩	10					1			**4**	2	3
ヨット	10			2					1	**6**	
家	10					1			2	1	**6**
分類再現率		0.9	0.5	0.3	0.2	0.5	0.6	0.6	0.4	0.6	0.6
分類適合率		0.41	0.56	0.38	0.5	0.56	0.60	0.75	0.50	0.66	0.5
F-値		0.56	0.53	0.34	0.28	0.52	0.60	0.66	0.44	0.62	0.54

表４．前景を切り出した画像の分類結果

人手による分類		k-means 法により得られたクラス		
テストデータ		クラス１	クラス２	クラス３
ジャンル	種類	花	象	トラ
植物	花	**26**	4	
動物	象		**30**	
人工画像 (イラスト)	トラ		17	**13**
分類適合率		1.00	0.59	1.00
分類再現率		0.86	1.00	0.44
F-値		0.93	0.74	0.62

次に、画像処理で重要なテクスチャ解析について検討した。従来法の場合、テクスチャ分類と学習の２ステップで構成される方式がよくみられる。本研究では、テキスト圧縮法を利用して、画像中に現れる本質的な頻出特徴パターンを発見することにより、学習ステップを不要とした。

実験では、Brodatz データベースと物体のテクスチャ表面を用いて、画像の回転、ズームなどの変化の下で認識能力の評価を行った。また、テクスチャ画像の検索についても実験を行った。提案法がかなり広範な変動に対してロバストであることを確認した。

（３）マルチメディアデータ解析：テキストや画像が同時出現することが多いウェブページの分類では、例えば、ページを決定木アルゴリズムを用いて画像、テキストの２つ

のカテゴリに分け、テキスト部位では単語間の繋がりを考慮する CNN-like word net などの特徴抽出法を使用し、画像部位に関しては領域輪郭線から領域の意味を推定する前段処理のあと、ベイズ推定法を用いてテキスト及び画像を結合させた分類を行っている。このような現在の主流アプローチでは、情報の種類毎に別々の手法や表現形式を用いたあと、面倒な重みパラメータの調整を要する統計的情報統合の段階が必要となる。この問題に対して、上記（１）のテキスト解析、（２）の画像解析研究の結果から本方式では双方を圧縮率特徴ベクトルで表現し、双方を結合したベクトルによって、画像とテキストの共起事象を表現する方式が有効であると考えられる。これによって情報統合の段階が単なるベクトル結合によって実現できる。関連データ収集の困難さから、実証については将来課題となった。

（４）ウェブプロトタイピング

インターネットの上クローラ（情報収集ソフトウェアロボット）を用いて、ウェブドキュメントを収集し、上記の（１）を用いたウェブページ分類実験を行った。

英語、日本語、中国語及びそれらの混在文書の分類が従来方式(Bag-of-words, N-gram)よりも良好な分類性能となることを確認した。

（５）その他

本研究の原理としている圧縮性に基づくデータ分類法（PRDC）に関し、Romania 政府機関主催の GEOSS ワークショップに招待され、画像データマイニングにおける課題解決のための圧縮性特徴量を利用したアプローチについて講演した。(その他①)

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計９件）

① Nuo Zhang and Toshinori Watanabe, Image Representation and Classification based on Data Compression, Proceedings of the 25th Annual ACM, pp.981-982, Mar. 22-26, 2010, Sierre, Switzerland.
② Nuo Zhang and Toshinori Watanabe, Documents Representation Based on Independent Compressibility Feature Space, Proceedings of ICAART'10, INSTICC, pp.217-222, Jan. 22-24, 2010, Valencia, Spain.
③ Nuo Zhang and Toshinori Watanabe, Documents Clustering Based on Optimized Compressibility Vector Space, Proceedings of International Conference on Computational Intelligence and Software Engineering (CiSE), IEEE, Dec. 11-13, 2009, Wuhan, China.
④ Toshinori Watanabe, On the Possibility of Highly Automated Image Information Mining: Problems and Possible Solutions, Workshop on Innovative Data Mining Techniques in Support of GEOSS, Aug. 31 - Sep. 2, 2009, Sinaia, Romania.
⑤ 小林 郁弥, 渡辺 俊典, 古賀 久志, ユークリッド空間内の点分布の外郭を求めるアルゴリズム, 信学技報 PRMU2009-171, pp.115-120, 2010/1/21, 京都.
⑥ Nuo Zhang, Daisuke Matsuzaki, Toshinori Watanabe and Hisashi Koga, Document Relation Analysis Based on Compressibility Vector, Proceedings of ICAART09, INSTICC, pp.255-260, Jan. 19-21, 2009, Porto, Portugal.
⑦ 平井 敦之, 張 諾, 渡辺 俊典, 古賀 久志, テキスト化を介した画像分類手法の提案, AI2008-42, pp.7-12, 2009/1/16, 東京.
⑧ Nuo Zhang, Toshinori Watanabe, Daisuke Matsuzaki and Hisashi Koga, A Novel Document Analysis Method Using Compressibility Vector, Proceedings of the 1st International Symposium on Data, Privacy, and E-Commerce (ISDPE'07), IEEE-CS, pp.38-40, Nov. 1-3, 2007, Chengdu, China.
⑨ 山崎 啓介, 張 諾, 渡辺 俊典, 古賀 久志, 高次元圧縮空間の対話的手法による次元縮小, 情報処理学会研究報告 2007-NL-181, pp.35-40, 2007/09/25, 東京.

〔その他〕

ホームページ等

①

http://events.rosa-rc.ro/docs/prezentations/sinaia%2031_08/ROSA_WS_T.Watanabe.pdf

６．研究組織

(1)研究代表者

渡邊 俊典 (WATANABE TOSHINORI)

電気通信大学・大学院情報システム学研究科・教授

研究者番号：10242348

(2)研究分担者
古賀　久志　（KOGA HISASHI）
電気通信大学・大学院情報システム学研究科・准教授
研究者番号：40361836

張　諾　（ZHANG NUO）
電気通信大学・大学院情報システム学研究科・助教
研究者番号：20436736

横山　貴紀　（YOKOYAMA TAKANORI）
電気通信大学・大学院情報システム学研究科・助教
研究者番号：10401621

(3)連携研究者