研究論文

アーカイブズ関連学会の論文が扱う研究分野の差異の可視化と論文投稿先予測モデルの作成

Visualizing Research Area Differences in Archives-Related Academic Societies and Developing a Model to Predict the Appropriate Journal for Paper Submission

新原俊樹[1*], 甲斐尚人[2], 小柏香穂理[3], 船越幸夫[4]
Toshiki SHIMBARU[1*], Naoto KAI[2], Kahori OGASHIWA[3], Sachio FUNAKOSHI[4]

1 西南学院大学情報処理センター
Information Processing Center, Seinan Gakuin University
〒814-8511 福岡県福岡市早良区西新 6-2-92
E-mail: shimbaru@seinan-gu.ac.jp
2 大阪大学附属図書館
University Library (Research and Development Unit), Osaka University
〒560-0043 大阪府豊中市待兼山町 1-4
3 お茶の水女子大学教学 IR・教育開発・学修支援センター
Center for Institutional Research, Educational Development, and Learning Support, Ochanomizu University
〒112-8610 東京都文京区大塚 2-1-1
4 記録管理学会事務局
The Records Management Society of Japan
〒267-0066 千葉県千葉市緑区あすみが丘 5-54-8
* 連絡先著者 Corresponding Author

新しい研究成果の公表作業の効率化を図るため，論文投稿先の候補となる学会誌の論文タイトルから内容を推測し，各学会が注力する研究分野を特定するとともに，最適な投稿先の予測モデルを作成した．まず，ChatGPT を活用してアーカイブズに関連する A~D の 4 つの学会誌に掲載された各論文のタイトルから内容を推測し，推測結果を集計して解析用データを作成した．このデータを次元削減して各論文の主成分軸上の分布を可視化し，クラスター分析によって 4 つの群を抽出した．学会誌別に各群の論文の収録編数を集計すると，A,B,D の各誌はそれぞれ一つの群の論文が突出して掲載されていたのに対し，C 学会誌は各群の論文を比較的偏りなく掲載する傾向がみられた．さらに，各論文の内容を説明変数，掲載先を目的変数とする学習データを複数の機械学習の手法で

学習させ，最適な投稿先の予測モデルを作成したところ，予測精度は AUC（Area Under the Curve）にして 0.88 前後の高い値が得られた.

To efficiently determine the best journal for publishing a new paper, this study developed a model that predicts the most suitable journal by inferring the content from each article's title. Initially, the study utilized ChatGPT to infer the contents from the titles of articles published in four different journals. The results were used to visualize the differences between articles, identifying four distinct clusters. It was observed that journals A, B, and D predominantly published papers from one particular cluster, whereas journal C demonstrated a more balanced approach, publishing papers from each cluster. Subsequently, the authors employed machine learning techniques to create a predictive model for identifying the optimal journal for submission. This model used the content of each paper as the explanatory variable and the choice of a journal as the dependent variable. The model achieved a high predictive accuracy with an AUC (Area Under the Curve) score of approximately 0.88.

キーワード：対話型生成AI，ChatGPT，論文投稿先予測モデル，主成分分析，クラスター分析

Keywords：Interactive Generative AI, ChatGPT, Journal Submission Prediction Model, Principal Component Analysis, Clustering Analysis

1 はじめに

学術的な研究成果が得られた場合に，それらを学術雑誌等に掲載することで，同じ研究分野に携わる研究者の間に広く共有していくことが重要である．このとき，効果的に研究成果を波及させるためには，研究内容によく適合した分野を対象とする学会の学術雑誌（以下，「学会誌」）を投稿先として的確に選定することが肝要である．ただし，そのためには，候補となる学会誌が過去にどのような論文を掲載してきたか把握し，投稿先に相応しいかどうか，複数の学会誌を比較しながら判断しなければならない．このとき，機械的な手法で個々の論文の内容を推測し，各学会の研究分野の違いを把握することができれば，一連の検討の効率化に大きく貢献し得る．

本研究では，図書館情報学及び人文社会情報学の一領域である記録管理やアーカイブズを研究対象とする 4 つの学会に注目し，これらの学会誌に掲載された個々の論文の内容を推測するとともに，各論文の研究の位置関係を可視化する．そのうえで，各学会の研究分野の差異を明らかにする．また，より直接的には，新規に執筆した論文の内容を踏まえて投稿先の助言が得られる仕組みが望ましい．そこで，各論文の研究内容と掲載先を学習データとして，新たな論文の内容に基づき最適な投稿先を予測するための様々な機械学習の手法の予測精度を比較し，最適な予測モデルの作成を目指す．

論文投稿先の推薦システムは既に大手出版社等から提供されているが，Elsevier[1]や Wiley[2]などの出版者によるものは，論文の採択率や掲載に係る費用など詳細な情報が得られる一方，検索の対象が当該出版者の発行誌に限られるなどの制約がある．また，The Bio Semantics Group[3] や Research Square[4]などの第三者によるサービスであれば，出版者の垣根を超えた横断的な検索が

可能である. こうした論文投稿先の推薦システムに関する近年の研究事例[5]~[8]を表 1 に整理した. [5]~[7]はコンピュータ・サイエンスや生物医学・生命科学など広範な分野にわたる多数の学術雑誌の中から最適な投稿先を予測することを目指している. このため, 分析対象は論文の要旨を含み, これらの情報から説明変数となるベクトルを算出する際には, TF-IDF（Term Frequency - Inverse Document Frequency）や Word2Vec などの自然言語処理の手法を用いている. また, 予測する投稿先の候補は多岐にわたるため, 最適な一誌（Top1）を予測した場合の精度は, 正解率にして 35~62%程度にとどまる.

一方, 研究分野の範囲が狭く, 投稿先の候補もある程度まで絞られているケースでは, [5]~[7]よりも簡易な手法で同程度の予測精度を実現できる可能性がある. 限定的な分野を対象とした[8]では, 194 編の各論文の特徴を表す 26 項目の説明変数と, ジャーナル・インパクトファクター（Journal Impact Factor™）に基づく 4 段階のランクを目的変数にしたデータから機械学習による予測モデルを作成し, AUC（Area Under the Curve）にして 0.85 程度の予測精度を得た. 本研究が着目したアーカイブズ分野の範囲やデータセットのサイズは[8]に近く, 予測モデルの精度を比較するうえで参考となる. ただし, [8]では, 論文の全文を逐一確認して説明変数を作成しているため効率的でない. そこで本研究は, 説明変数の作成に係る作業を効率化するとともに予測精度の高度化を図るため, 論文の内容を端的に示したタイトルに着目し, 説明変数も生成 AI を用いた機械的な手法を採用することとした.

2 論文一覧の取得

本研究では, 記録管理やアーカイブズを研究分野とする 4 つの学会を対象とした. 各学会の設立時期は, 記録管理学会とアート・ドキュメンテーション学会が 1989 年, 日本アーカイブズ学会が 2004 年, デジタルアーカイブ学会が 2017 年である. これらの学会が刊行した過去約 20 年間の学会誌に掲載された合計 239 編の論文のタイトルを収集した. 各学会誌のタイトルの収集状況を表 2 に示す.

ここで収集対象としたのは, 査読付の論文（原著論文, 及び, これに準じた研究ノート）とし, 全国大会や研究会の予稿は対象外とした. 査読を通じて各学会の編集委員会の承認を得た論文に限定することで, 各学会が研究の対象とする分野を明確にすることができる. なお, 以降は表 2 の最左列のとおり, 各学会を A~D 学会と呼ぶ.

表 1 論文投稿先の推薦システムに関する近年の研究事例

先行研究	分野	学会数（分類クラス数）	学習データとした論文数	分析対象	説明変数（ベクトル）の作成手法	予測精度	
						正解率	AUC
Wang et al. (2018) [5]	コンピュータ・サイエンス	28学会+38会議	14,012（2013~2014年）	要旨	・TF-IDF ・カイ二乗特徴選択	35.03%（Top1） 61.37%（Top3）	0.94
Feng et al. (2019) [6]	生物医学・生命科学	1,130誌	880,165（2007~2016年）	要旨	・word2vec + CNN	39%（Top1） 61%（Top3） 68%（Top5） 76%（Top10）	–
Huynh et al. (2022) [7]	コンピュータ・サイエンス	（不明）	414,512	タイトル,要旨,キーワード + 各誌の「目的と範囲」	・DistilBERT + CNN1D	62.46%（Top1） 90.32%（Top3） 94.89%（Top5） 97.96%（Top10）	–
手塚ほか (2021) [8]	肺がん領域臨床研究	4クラス	194	全文	26項目それぞれに適合するか否か判定	–	0.86

表 2 各学会誌の論文タイトルの収集状況

	学会名（学会誌名）	設立年	巻･号（期間）	タイトル数
A	記録管理学会 （レコード・マネジメント）	1989	48～84号[9] （2004年10月～2023年3月）	99
B	アート・ドキュメンテーション学会 （アート・ドキュメンテーション研究）	1989	9～30号[10] （2001年3月～2022年5月）	80
C	日本アーカイブズ学会 （アーカイブズ学研究）	2004	3～36号[11] （2005年11月～2022年6月）	39
D	デジタルアーカイブ学会 （デジタルアーカイブ学会誌）	2017	2～7号[12] （2018年1月～2023年3月）	21

3 解析用データの作成

次に，各論文のタイトルから研究内容を推測し，解析に用いるデータを作成する．このとき用いる自然言語処理の手法として，例えば，BERT（Bidirectional Encoder Representations from Transformers）[13]，Sentence BERT[14]，fastText[15]などにより，単語や文章を，それらの意味の特徴を表現したベクトル（分散表現）に変換して演算する手法が考えられるが，本研究ではプログラミングによらず手軽に扱える手法として，対話型生成 AI の GPT-4（Generative Pre-trained Transformer 4）[16]を用いたChatGPT[17]を活用して研究内容の推測を行った．

ChatGPT は，OpenAI 社が 2022 年 11 月に公開した大規模言語モデルに基づく生成 AI のサービスであり，自然言語処理を専門としない者でも手軽に大規模言語モデルを用いて自然言語処理の手法で問題の解決に取り組むことが可能になった．その具体的な活用方法についても，数多くの提案がなされている．

ChatGPTを用いた推測手順は，新原俊樹らの報告[18]に倣った．具体的には，ChatGPT に対して，論文のタイトル一覧とA~Lの12種類の研究の特徴（以下，「観点」）を表現した複数の単語（後述の表3に記載）を提示し，個々の論文がそれぞれの単語に関係するのか否か，1又は0で判定させ，表形式で回答させる指示（以下，「プロンプト」）を与えた．

このプロンプトでは，単にタイトル内に特定の単語が含まれているか否かに限らず，単語の意味も踏まえた比較を行わせている．これにより，例えば「大学」の語を含むタイトルは観点 I（教育,学校）に，「アメリカ」や「韓国」の語を含むタイトルは観点 J（外国,国際標準）に関係すると判定される．

ただし，ChatGPT の回答にはゆらぎがあるため，同一のタイトルを対象として同じプロンプトを与えても回答が異なることがある．API（Application Programming Interface）経由で直接 GPT-4 を動作させる場合は，パラメータを設定することで回答の再現性を高める工夫も可能である．しかし，ChatGPT ではこれらのパラメータの設定ができない．そこで，回答のゆらぎが結果に与える影響を減らすため，同一のタイトルを対象とした同じプロンプトを 10 回繰り返し与え，ChatGPT が「関係あり」と判定した回数（すなわち，1~10 の値）を観点別に集計し，これを解析用データとした．なお，同じプロンプトを繰り返し与える際，各回の回答がそれ以前の回答の影響を受けないように，プロンプトの入力履歴を ChatGPT に学習させない設定とした．

得られた回答の全要素（239 編×12 観点=2,868 要素）について，各要素の値の度数分

布(図 1)を見ると, 0 又は 10 の値をとる要素(10 回のうち, すべて関係なし又は関係ありと判定された要素)は2,531に上り, 全体の88 %を占めた. このように, 繰り返し同じプロンプトから得た回答を積算し集計することで, 作成するデータの再現性を高めた. 本研究では先行研究[18]と同じ論文一覧を用いたため, 作成した解析用データも同じものになった.

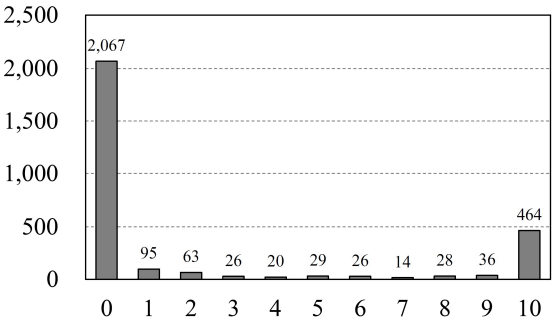

図 1 解析データの各要素の値の度数分布

4 主成分分析による次元削減と各主成分の解釈

作成した解析用データには, 各論文の研究内容を特徴づける説明変数となる 12 種類の観点が含まれている. そこで, 主成分分析による次元削減を行い, 各論文の研究内容の位置関係を可視化しやすい主成分を抽出した. なお, 分析に当たっては, Python 用の機械学習ライブラリである scikit-learn[19]を使用し, 分析環境は Google 社の Colaboratory を利用した[20].

主成分分析の結果として抽出した第 1 主成分(PC1)～第 5 主成分(PC5)の固有値, 寄与率, 累積寄与率, 主成分に対する各観点の主成分負荷量を表 3 に示す. 固有値は, 各主成分が含む元の情報量の大きさを表し, これを元の情報量全体に占める割合に変換したものが寄与率である. 各主成分の寄与率と累積寄与率を見ると, 第 5 主成分までの累積寄与率は 0.589 となり, この 5 つの主成分で元の情報量全体の約 60 %を保持していることがわかる. また, 主成分負荷量は各観点と主成分の間の相関の度合いを表し, 主成分負荷量の絶対値が大きい観点ほど主成分と強い相関がある. 表 3 の結果では, 第 1 主成分は B,C,G, 第 2 主成分は F,K,L, 第 3 主成分は I,J, 第 4 主成分は D, 第 5 主成分は E,H とそれぞれ相関が高くなっている(表 3 中の主成分負荷量の絶対値が 0.5 以上の値を網掛けで表示).

この結果を踏まえて, 各主成分軸上に位置する論文の内容を解釈すると, 第 1 主成分の軸上では, 公的機関の文書等の整理や分類を扱う論文ほどプラス側に, 電子又はデジタルデータを対象としたものほどマイナス側に位置している. この関係は,「公的機関の文書を扱う論文ほど, 電子又はデジタルデータに言及したものが少ない」「電子データ・デジタルデータを扱う論文の中には, 公的機関の文書を対象としたものが少ない」ことを意味している. 公刊された論文にこうした特徴がみられる背景として, 行政機関をはじめとする公的機関において記録や資料の電子化・デジタル化が進んでいない実状を反映している可能性がある. また, 第 2 主成分上では, 図書館や博物館, 美術館を題材にした研究がプラス側に, アーカイブに関するものがマイナス側に寄っている. ただし, 本研究の対象がすべてアーカイブズ関連の学会誌に掲載された論文であることを考慮すると, 両者は相反する関係ではなく, アーカイブズの中でも代表的な図書館, 博物館, 美術館を題材とした論文群がプラス側に, それ以外のものがマイナス側に位置付けられているものと考えられる. 第 3 主成分上では, 教育・学校関係の論文がプラス側, 諸外国の制度や事例を扱った論文がマイナス

表 3 主成分分析の結果

主成分		PC1	PC2	PC3	PC4	PC5
固有値		2.026	1.487	1.299	1.172	1.082
主成分負荷量	A(近世,近代,戦前,戦中,戦後)	0.268	0.068	-0.294	-0.470	0.235
	B(電子,デジタル,情報,データ)	-0.685	-0.038	-0.025	-0.228	-0.122
	C(政府,行政機関,自治体,行政文書,公文書,公文書館)	0.672	0.079	0.018	-0.338	-0.106
	D(ビジネス,企業,民間)	0.067	-0.260	-0.331	0.574	0.313
	E(写真,画像,映像,動画)	-0.453	-0.140	0.068	-0.456	0.541
	F(アーカイブ)	-0.258	-0.716	-0.166	0.125	-0.111
	G(記録管理,資料整理,文書分類,ﾌｧｲﾘﾝｸﾞｼｽﾃﾑ)	0.717	0.084	0.231	0.133	0.084
	H(事例,調査)	0.010	0.082	0.352	0.337	0.542
	I(教育,学校)	-0.044	-0.349	0.638	0.047	-0.436
	J(外国,国際標準)	0.144	-0.045	-0.695	0.092	-0.255
	K(図書館,博物館,文化財)	-0.300	0.660	-0.026	0.066	-0.203
	L(芸術館,絵画,芸術,アート)	-0.362	0.550	-0.033	0.295	-0.028
寄与率		0.169	0.124	0.108	0.098	0.090
累積寄与率		0.169	0.293	0.401	0.499	0.589

側に多く分布する. この関係は,「教育・学校のテーマを扱った論文の中に, 諸外国の事例に言及したものが少ない」「諸外国の制度や事例に関する論文の中に教育・学校に着目したものが少ない」ことを意味する. 実際に解析用データを確認すると, ChatGPT による 10 回すべての回答で「教育・学校に関係あり」とされた論文は 17 編,「諸外国の制度・事例に関係あり」とされたものは 47 編あったが, 両者に共通したのはわずか 1 編であった. さらに, 第 4 主成分上では, 企業や民間の団体を対象としたビジネス・アーカイブズに関する論文がプラス側に, 第 5 主成分上では, 画像や映像, 事例調査に関するものがプラス側に位置した.

5 各論文の位置関係の可視化

続いて, 次元削減後の各論文の主成分得点をもとに, これらの位置関係を可視化した(図 2). 図 2 は各主成分間の散布図である. 各論文の位置を学会誌別に色分けして表示し, 学会誌別に算出した論文の分布の重心位置を☆印で示した. 第 1 主成分軸上(図 2(a),(b),(c),(d)の横軸上)から見ると, A 学会と C 学会, B 学会と D 学会の重心がそれぞれ近い. 一方, 第 2 主成分軸上(図 2(a)の縦軸上, 及び, 図 2(e),(f),(g)の横軸上)から見ると, B 学会の重心のみが他の 3 つの学会からやや離れている. 第 3, 第 4, 第 5 主成分軸上では, いずれの重心も近い位置にある. このように, どの主成分軸上から見ても A 学会と C 学会の重心が近い位置にあることがわかる.

ここで, 各論文の主成分得点をもとに k-means 法でクラスター分析を行い, 本研究が対象とした学会の数に合わせて 1~4 群のクラスターを抽出した(図 2 中に灰色の枠線で各群を表示). 図中の各群の位置関係から, 主に公的機関の文書や資料の整理・分類を扱ったものが 1 群, これに対し, 電子又はデジタルデータを扱った研究のうち, 特に図書館・博物館・美術館を対象にしたものが 2 群, その他のデジタルアーカイブに関連したものが 3 群, 諸外国の制度や事例又はビジネス・アーカイブズに関するものが 4 群だと言える.

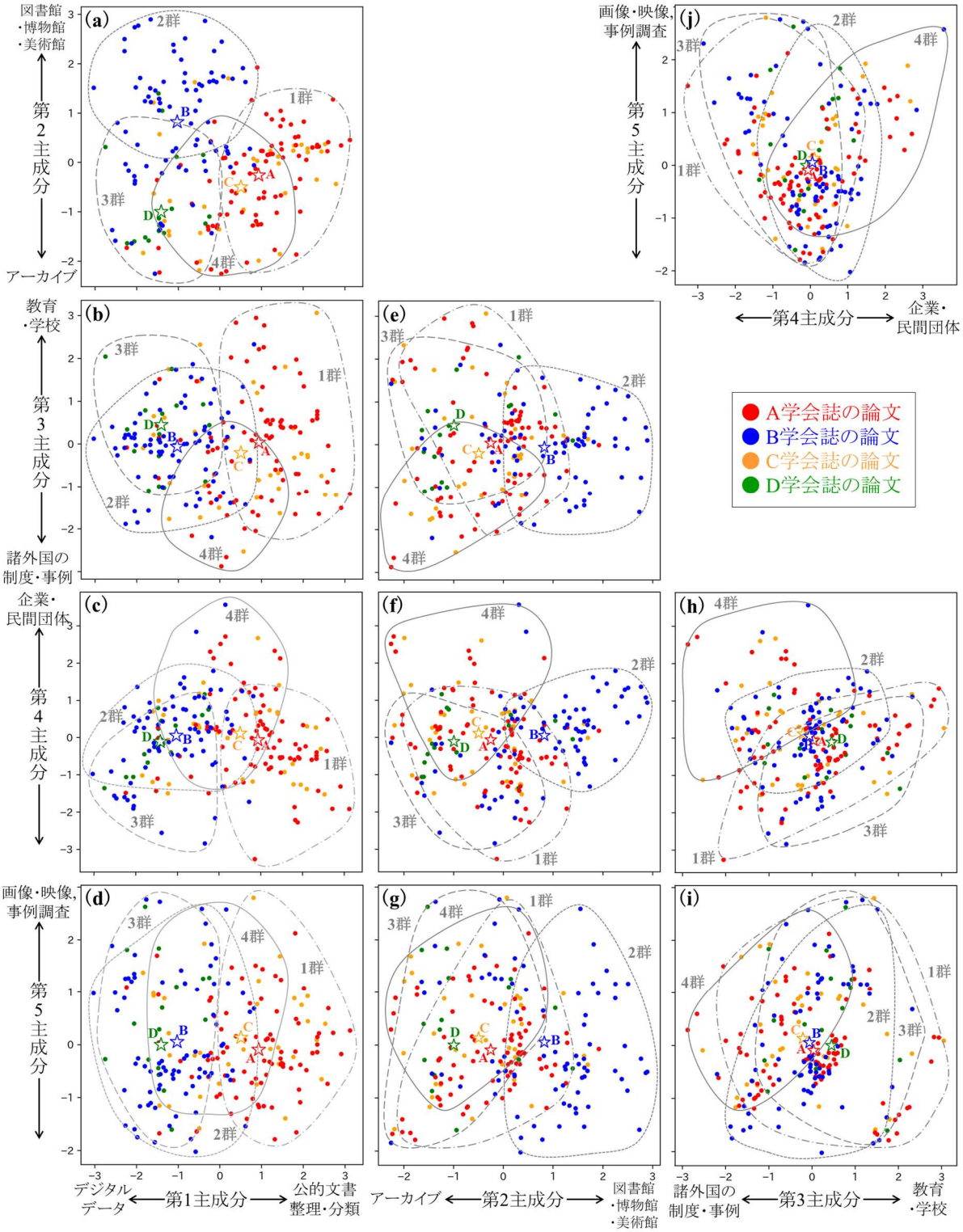

図 2　各論文の主成分得点の分布

表 4 学会誌別 各群の論文数と全体に占める割合

	1群	2群	3群	4群	合計
A学会	64(65%)	4(4%)	7(7%)	24(24%)	99(100%)
B学会	3(4%)	52(65%)	20(25%)	5(6%)	80(100%)
C学会	13(33%)	4(10%)	8(21%)	14(36%)	39(100%)
D学会	0(0%)	2(10%)	16(76%)	3(14%)	21(100%)

次に，学会誌別に 1~4 群のどの論文が多く含まれるか集計した（表 4）．表 4 を見ると，A 学会誌に 1 群，B 学会誌に 2 群，D 学会誌に 3 群の論文が多く掲載されている．この傾向は，各学会が設立趣意書などで表明している研究対象領域（下記の下線部）とも整合している．

「〔筆者注：A 学会が扱う『記録管理』について〕企業・行政機関などの組織が活動を行う上で，さまざまな記録（文書など）が作成されます．記録管理とはその記録の作成から組織化，活用，保管，最終処置（永久〈アーカイブ〉保存・廃棄）までのライフサイクルを対象とした統合的な管理をいいます．」[21]

「〔筆者注：B 学会の学術的背景について〕本学会には，図書館司書，学芸員，アーキヴィスト，情報科学研究者，美術史・文学史・音楽史・メディア史・文化史・自然史研究者など，約 350 名の正会員，学生会員，賛助会員が所属しています．従来の美術館／博物館・図書館・公文書館・アーカイヴおよび学会といった機関や職能を超領域的に融合する新しい学術団体として，本学会は，新しい未知な課題に取り組む方々の参加をえて，活動を展開しています．」[22]

「〔筆者注：D 学会誌の論文掲載方針について〕本会誌はデジタルアーカイブに関連し，またその発展に貢献する記事を掲載し，以てデジタルアーカイブの振興に寄与する．分野としては，デジタルアーカイブの理論，提言，実践，情報／知識の収集，整理，蓄積，検索および各種解析，利用などに関するもの，などである．」[23]

一方，C 学会誌には 4 群の論文が多いものの，1 群との差は僅かであり，他の 2 群や 3 群の論文が占める割合も小さくない．この傾向は，C 学会が注目する研究領域（下記の下線部）が他の 3 学会と比較してやや広く解釈し得ることも一因だと考えられる．

「〔筆者注：C 学会が扱う『アーカイブズ』の研究領域について〕このアーカイブズに関する科学的研究は，(1)アーカイブズの管理に関する研究，(2)アーカイブズの成立・構造・伝来などに関する研究，(3)アーカイブズの教育・普及に関する研究などから構成されており，歴史学，社会学，情報学など既存の様々な学問分野の学理と連携しつつ，独自な領域をもつものである．この科学的研究は，アーカイブズの保存及び関連する諸課題の解決に資するという役割を担うものでもある．」[24]

6 各学会誌の論文の特徴の変化

これまで，各学会誌に掲載された論文の特徴の差異を明らかにしてきたが，同じ学会誌においても過去 20 年間に掲載された論文の特徴が変化した可能性もある．そこで，長期にわたり刊行されてきた A,B,C の各誌に掲載された論文を 2013 年以前（以下，「前期」）と 2014 年以降（以下，「後期」）の 2 つの期間に分け，解析用データから，A~L の各観点について ChatGPT に「関係あり」と判定された論文の数を集計した（表 5）．また，各期間の論文のうち各観点に関係する論文が占める比率を求めて前後期で比較し，フィッシャーの正確確率検定（片側検定，有意水準 5%）に基づき比率が大きい方に*印を付した．

表 5 各観点に「関係する」と判定された論文数

学会誌	期間	論文数	A	B	C	D	E	F	G	H	I	J	K	L
A学会	2013年以前	47	1	6	14	6*	3	13	27*	2	6	5	2	0
	2014年以後	52	7*	7	20	1	2	12	18	8	6	17*	1	1
B学会	2013年以前	54	7	22	1	2	8	5	6	5	3	12*	14	20
	2014年以後	26	0	13	0	0	6	7*	1	7*	1	1	4	8
C学会	2013年以前	20	1	4	3	3	2	9	5	3	4	7	1	0
	2014年以後	19	3	1	4	2	2	6	9	3	1	5	2	0

この結果を見ると，A 学会誌で後期に掲載された論文は，前期と比較して観点 D（ビジネス,企業,民間）や観点 G（記録管理,資料整理,文書分類,ファイリングシステム）に関する論文が減った一方，観点 A（近世,近代,戦前,戦中,戦後）や観点 J（外国,国際標準）に関するものが増えた．また，B 学会誌では前期から後期にかけて観点 J の論文が減ったが，観点 F（アーカイブ）や観点 H（事例,調査）に関係したものが増えていた．C 学会誌については，前後期を通じて有意な変化は確認できなかった．

7 論文投稿先予測モデルの作成

より直接的なアプローチとして，新規に執筆した論文のタイトルからその内容を推測し，最適な投稿先を提示できる予測モデルの作成を目指す．このモデルの作成に当たり，先の解析用データに含まれる 12 の観点を説明変数，投稿先の学会誌を目的変数とする 239 編のデータを用いた．このデータをランダムに学習データ（70 %, 167 編）とテストデータ（30 %, 72 編）に分割し，さらに学習データをランダムに 10 分割したうえで，そのうち 1 つを検証データとして入れ替えながら 15 種類の機械学習の手法による予測精度を比較した．なお，一連の検証において，Python の機械学習ライブラリである PyCaret[25]を使用した．

検証の結果，予測精度が上位であった 4 つのモデルについて，ランダムサーチによってハイパーパラメータを最適化した後の予測精度を求めた（表 6）．これらの精度は，多数の雑誌の中から最適な一誌（Top1）を提案する先行研究[5]~[7]の精度よりも高く，限定的な分野内で投稿先を予測した[8]の精度と同程度かやや上回るものであった．このように，論文のタイトルのみを使用し，生成 AI を用いた機械的な手法で効率的に説明変数を作成できる点に本研究の手法の強みがある．

次に，これらのモデルを用いて，テストデータ（72 編）の掲載先を予測した結果（混同行列）を表 7 に示す．表 7 を見ると，いずれのモデルも C 学会誌に掲載されている論文を誤って A 学会誌と予測した例が多いことがわかる．これは，図 2 のいずれの主成分軸上から見ても A,C 各学会誌の論文の重心が近く，また，表 4 において両誌が含む各群の論文の割合が近いことからも推察されるように，両誌に掲載される論文の研究分野が近いことに起因していると考えられる．

表 6 予測精度上位 4 モデルの AUC と正解率

モデル名	AUC	正解率（Accuracy）
Extra Trees Classifier	0.881	0.707
Light Gradient Boosting Machine	0.874	0.725
Extreme Gradient Boosting	0.873	0.732
Random Forest Classifier	0.869	0.714

表 7 予測精度上位 4 モデルによるテストデータ(72 編)の掲載先の予測結果(混同行列)

Extra Trees Classifier

		予測結果			
		A	B	C	D
実際の掲載誌	A	23	2	3	2
	B	1	20	0	3
	C	6	4	2	0
	D	0	2	1	3

Extreme Gradient Boosting

		予測結果			
		A	B	C	D
実際の掲載誌	A	25	4	0	1
	B	1	22	0	1
	C	7	4	1	0
	D	0	4	1	1

Light Gradient Boosting Machine

		予測結果			
		A	B	C	D
実際の掲載誌	A	23	3	2	2
	B	2	19	0	3
	C	6	4	2	0
	D	0	3	1	2

Random Forest Classifier

		予測結果			
		A	B	C	D
実際の掲載誌	A	23	3	2	2
	B	2	20	0	2
	C	8	4	0	0
	D	1	2	0	3

8 まとめ

研究成果の内容に合わせて最適な投稿先の学会誌を選定するための検討の効率化を目的として, 本研究ではまず, 対話型生成 AI の ChatGPT を活用し, 4 つの学会誌に掲載された各論文のタイトルから内容を推測して解析用のデータを作成した.

次に, 解析用データを次元削減し, 各論文の研究内容の違いを可視化しやすい主成分を抽出して, 各主成分軸の持つ意味について考察した. 論文の特徴の分散が最大となる第 1 主成分上では, 公的機関の文書を扱う論文ほどプラス側に, 電子又はデジタルデータに言及した論文ほどマイナス側に分布することが示された. この背景として, 政府や地方公共団体をはじめとする公的機関で記録や資料の電子化・デジタル化が進んでいない実状が反映されている可能性がある. 第 2 主成分上では, 図書館や博物館, 美術館を題材にした研究がプラス側に, アーカイブに関するものがマイナス側に寄っていた. ただし, 両者は相反する関係ではなく, アーカイブズの中でも代表的な図書館, 博物館, 美術館を題材とした論文群とそれ以外の論文を分ける軸になっているものと考えられる. 第 3 主成分上では, 教育・学校関係の論文がプラス側, 諸外国の制度や事例を扱った論文がマイナス側に多く分布した. 実際に解析用データを確認したところ,「教育・学校」と「諸外国の制度・事例」の両方に関係する論文はわずかであった.

続いて, 各論文の主成分得点に基づく分布を可視化するとともに, クラスター分析によって 4 つの群を抽出した. 学会誌別に各群の論文の収録編数を集計したところ, A 学会誌は 1 群(主に公的機関の文書や資料の整理・分類を扱った論文), B 学会誌は 2 群(特に図書館・博物館・美術館を対象にした論文), D 学会誌は 3 群(その他のデジタルアーカイブに関連した論文)が最も多く, 各学会が設立趣意書などを通じて表明している研究対象領域とも整合していた. 一方, C 学会誌は 1 群と 4 群(諸外国の制度や事例又はビジネス・アーカイブズに関する論文)を中心としつつも, 比較的偏りなく各群の論文を掲載しており, 他の 3 学会と比べて研究分野の幅広さが窺えた.

さらに，より直接的な手法として，各論文の研究内容と掲載先を学習データとし，新規の論文の内容に基づき最適な投稿先を予測するための様々な機械学習の手法を比較して最適な予測モデルを作成したところ，その精度は AUC にして 0.88 前後となり，先行研究の予測モデルと同程度以上の結果が得られた．今後，同じ手法で他の研究分野でも同様の成果が得られるか検証を続ける価値がある．

9 データ利用可能性宣言

本研究中に作成された，分析に用いたデータセットは，妥当な理由があれば責任著者から入手可能である．

参考文献

［1］Elsevier: “Journal Finder,” https://journalfinder.elsevier.com（2024 年 4 月 1 日参照）．

［2］Wiley: “Find journals that match your manuscript,” https://journalfinder.wiley.com（2024 年 4 月 1 日参照）．

［3］The Bio Semantics Group: “Jane Journal /Author Name Estimator,” https://jane.biosemantics.org（2024 年 4 月 1 日参照）．

［4］Research Square: “Journal Guide,” https://www.journalguide.com（2024 年 4 月 1 日参照）．

［5］Wang, D.; Liang, Y.; Xu, D.; Feng, X.; Guan, R.: “A Content-based Recommender System for Computer Science Publications,” *Knowledge-Based Systems*, Vol. 157, pp. 1-9, 2018, https://doi.org/10.1016/j.knosys.2018.05.001

［6］Feng, X.; Zhang, H.; Ren, Y.; Shang, P.; Zhu, Y.; Liang, Y.; Guan, R.; Xu, D.: “The Deep Learning–Based Recommender System ‘Pubmender’ for Choosing a Biomedical Publication Venue: Development and Validation Study,” *Journal of Medical Internet Research*, Vol. 21, No. 5, e12957, 2019, https://www.jmir.org/2019/5/e12957

［7］Huynh, S. T.; Dang, N.; Nguyen, D. H.; Huynh, P. T.; Nguyen, B. T.: “FPSRS: a fusion approach for paper submission recommendation system,” *Applied Intelligence*, Vol. 53, pp. 8614-8630, 2022, https://doi.org/10.1007/s10489-022-04117-8

［8］手塚有佳里; 橋詰薫; 坂井貞興:「機械学習プラットフォームである DataRobot を用いた臨床研究結果の最適な論文投稿先予測モデルの生成」, 2021 年度人工知能学会全国大会論文集, 2Xin5-08, 2021.

［9］国立研究開発法人科学技術振興機構:「J-STAGE レコード・マネジメント」, https://www.jstage.jst.go.jp/browse/rmsj/list/-char/ja（2024 年 4 月 1 日参照）．

［10］アート･ドキュメンテーション学会:「アート・ドキュメンテーション研究」, http://www.jads.org/pub/kenkyu.html（2024 年 4 月 1 日参照）．

［11］日本アーカイブズ学会:「アーカイブズ学研究」, http://www.jsas.info/?cat=7（2024 年 4 月 1 日参照）．

［12］国立研究開発法人科学技術振興機構:「J-STAGE デジタルアーカイブズ学会誌」, https://www.jstage.jst.go.jp/browse/jsda/-char/ja（2024 年 4 月 1 日参照）．

［13］Devlin, J.; Chang, M.; Lee, K.; Toutanova, K.: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *arXiv*, 1810.04805, 2018, https://arxiv.org/abs/1810.04805

［14］Reimers, N.; Gurevych, I.: “Sentence-

BERT: Sentence Embeddings using Siamese BERT-Networks,” *arXiv*, 1908.10084, 2019, https://doi.org/10.48550/arXiv.1908.10084

［15］Joulin, A.; Grave, E.; Bojanowski, P.; Mikolov, T.: “Bag of Tricks for Efficient Text Classification,” *arXiv*, 1607.01759, 2016, https://doi.org/10.48550/arXiv.1607.01759

［16］OpenAI et al.: “GPT-4 Technical Report,” *arXiv*, 2303.08774, 2023, https://doi.org/10.48550/arXiv.2303.08774

［17］OpenAI: “Introducing ChatGPT,” https://openai.com/blog/chatgpt（2024 年 4 月 1 日参照）.

［18］新原俊樹; 甲斐尚人; 小柏香穂理; 船越幸夫:「ChatGPT を活用した研究データの作成事例」, 情報知識学会誌, Vol. 34, No. 1, pp. 18-23, 2024, https://doi.org/10.2964/jsik_2023_031

［19］Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.: “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, Vol. 12, No. 85, pp. 2825-2830, 2011.

［20］Bisong, E.: “Google Colaboratory,” *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Apress, Berkeley, CA.*, pp. 59-64, 2019.

［21］記録管理学会:「記録管理学会へのお誘い」, https://rmsj.smoosy.atlas.jp/ja/osasoi（2024 年 4 月 1 日参照）.

［22］アート･ドキュメンテーション学会:「趣旨」, http://www.jads.org/guide/guide.html（2024 年 4 月 1 日参照）.

［23］デジタルアーカイブ学会:「『デジタルアーカイブ学会誌』編集方針」, https://digitalarchivejapan.org/gakkaishi/toukou/aboutthisjournal/（2024 年 4 月 1 日参照）.

［24］日本アーカイブズ学会:「日本アーカイブズ学会会則」, http://www.jsas.info/?page_id=47（2024 年 4 月 1 日参照）.

［25］Moez, A.: “PyCaret,” https://pycaret.org（2024 年 4 月 1 日参照）.

（2023年12月18日　受付）
（2024年 4月24日　採択）
（2024年 6月 7日　J-STAGE早期公開）