レコード・マネジメント №85　pp. 54〜65（2023）

プロジェクト研究

機械学習を用いた行政文書の第一次選別機能の開発と性能評価

新原　俊樹[1]

［抄録］地方公共団体では、保存期間が満了した行政文書の一部を特定歴史公文書として選定する評価選別作業の効率化が課題となっている。本研究は、熊本県の知事部局において2016〜2020年度に保存期間満了を迎えた文書を対象として、過年度（2016〜2019年度）の文書の第一次選別結果に基づき、コサイン類似度に基づく文書の同一判定手法を用いて最新年度（2020年度）の文書の第一次選別を自動判定で行う手法を開発した。その結果、自動で判定を下すことができた文書数は全体の58％に上ったほか、自動判定結果と委員による実際の判定結果を比較したところ、再現率は60％、適合率は32％であった。各指標はトレードオフの関係にあり、全ての指標を共に改善することは容易でないが、判定精度の向上に寄与し得るものとして、（１）他の機械学習モデルとも比較し、最適なモデルの選定、（２）文書の内容に合わせて適切な名称を付与するしくみの導入、（３）過年度の第二次選別結果の第一次選別の過程への反映が今後の研究課題である。

［キーワード］行政文書、特定歴史公文書、機械学習、コサイン類似度

Development And Evaluation of an Automatic Judgment Method Using Machine Learning for the Primary Screening of Historical Public Records

Toshiki SHIMBARU

［Abstract］Local governments face the problem of improving the efficiency of the selection process for historical public records. This study developed an automatic judgment method for the primary screening of historical public records by machine learning, using the administrative documents of Kumamoto Prefecture as a case study. Automatic determination was possible for 58% of all documents. Comparing the automatic judgment results with the actual manual judgment showed a recall rate of 60% and a precision rate of 32%. However, each rate has a trade-off relationship, and improving each index is difficult. Efforts to improve the accuracy of judgments are the following: （1）Selection of the most appropriate machine learning model exceeding the cosine similarity, （2）Introduction of a system to assign appropriate names to documents according to their contents, and （3）Reflection of the secondary screening results in the past years in the primary screening process.

［Keywords］Administrative Documents, Historical Public Records, Machine Learning, Cosine Similarity

[1]しんばる　としき　西南学院大学情報処理センター
〒814-8511　福岡市早良区西新6-2-92
Toshiki Shimbaru（Information Processing Center, Seinan Gakuin University）
6-2-92 Nishijin, Sawara-ku, Fukuoka-City, Fukuoka 814-8511

（原稿受理　2023.9.15）

目次

1　はじめに

行政文書は、行政機関による政策などの意思決定プロセスを明示し、国民や市民への説明責任を果たすものである。行政機関は、誰もが行政文書にアクセスできるように適切な文書管理体制を構築することが不可欠である。我が国では2011年に公文書管理法[1)]が施行されて以来、行政機関は各文書に保存期間を定めて適切に管理し、保存期間満了後には必要なものを特定歴史公文書として公文書館に移管すること、移管された特定歴史公文書は同館において永久に保存することが義務付けられた。また、2018年には同法の関連規則が改正され、原則として各課単位で全ての文書を行政文書ファイル管理簿（以下、「管理簿」）に登録し、公表することとされた。こうした制度の変更により、現在、行政機関が保管する行政文書には「組織内部での共有・継承のしやすさ」だけでなく、組織の外から閲覧した際の「有無（何があるのか）」や「所在（どこにあるのか）」の分かりやすさも求められている。

一方、地方公共団体においても、同法第34条により文書を適正に管理するための努力義務が課せられている[2)]。もとより、県民や市民に密接に関わる組織の性質上、文書へのアクセスのしやすさは、国の機関に劣らず求められている。しかし、地方自治研究機構による2023年４月時点の調査[3)]によると、公文書管理条例を制定しているのは、都道府県が18/47（38％）[4)]、政令指定都市が7/20（35％）、市区町村が33/1,718（２％）、また宮間純一による2022年８月時点の調査[5)]によれば、公文書館を設置しているのは、都道府県が42/47（89％）、政令指定都市が10/20（50％）、市区町村が34/1,718（２％）にとどまっており、地方公共団体における文書管理体制の構築は進んでいない。その背景には「人材不足」と「財源不足」があるが、地方公共団体はこうした厳しい環境にあっても保存期間中において文書を適切に管理し、誰もが文書にアクセスできるしくみを整え、保存期間満了後も意思決定の過程について説明責任を果たすことが求められる。

1.1　問題の所在

こうした中、熊本県では、2011年から全国に先駆けて行政文書管理条例[6)]を制定し、公文書館を設置する代わりに各職員に責任を持たせ、文書の作成から廃棄又は永久保存に至るまでの文書のライフサイクルに基づく文書管理を行わせている[7)]。財源に依存せず、職員の手で文書管理制度の目的を達成することができれば、この体制がほかの地方公共団体にも波及し、地方行政における文書管理制度の発展が期待できる。一方で、これまでの先行研究[8)9)10)11)12)]により同県の文書管理体制に内在する複数の課題が指摘されている。例えば、同県では保存期間が満了した廃棄予定の文書について、第三者評価委員（以下、「委員」）が文書のタイトルを点検し、特定歴史公文書の判断基準となる「評価選別基準」に照らして廃棄が不適当だと考える場合は、廃棄の方針を保留する（以下、「第一次選別」）。第一次選別で「保留」とされた文書は、委員があらためて文書の現物を精査し、「廃棄」又は「永久保存」の判定をする（以下、「第二次選別」、また、これらの一連の選別作業をまとめて「評価選別」という）。しかし、毎年約５万件の文書が保存期間満了を迎えるた

め、現行の手法では委員が評価選別を終えるまでに数か月を要しており、持続可能な体制になっていない。評価選別の自動化・効率化が喫緊の課題である。

1.2　本研究の目的

緒方克治ら[10]はこの問題を解決するため、2020年度末に保存期間満了を迎えた文書の第一次選別結果に基づき、選別のルールを整理して実装した自動判定ツールを作成し、これまで１か月以上を要していた第一次選別を僅か30秒で達成するなどの成果を上げた。ただし、ツールによる判定結果は、委員が「保留」とした文書数の約２倍の文書を「保留」と判定する傾向にあるほか、自動判定ツールを別の年度の文書に適用すると判定精度が低下するなどの課題を残しており、現段階では実用化に至っていない。

そこで本研究では、機械学習の一つであるコサイン類似度による文書の同一判定手法を用いて、過去の第一次選別記録に基づき最新年度の文書の第一次選別を自動で行う手法を開発した。その上で、判定精度の評価を行い、第二次選別結果も取り入れた改善策について検討した。本手法を確立することで、地方行政機関が扱う文書の特徴を反映したボトムアップ型の文書管理体制が充実し、他の地方公共団体における文書管理体制の確立にも貢献できる。

2　第一次選別の自動判定手法の開発

2.1　対象とした行政文書

本研究では、熊本県の知事部局（広域本部や地域振興局を含む）において、2016～2020年度の間に保存期間満了を迎えた文書を対象とした。各年度の総文書数は、26,598件（2016年度）、37,424件（2017年度）、36,883件（2018年度）、36,296件（2019年度）、36,184件（2020年度）であり、2017年度以降は概ね同程度の数の文書が保存期間満了を迎えている。これら全ての文書に、委員による第一次選別の判定結果（「廃棄」又は「保留」のいずれかの値）が付されている。このうち、2016～2019年度（以下、「過年度」）の各文書の判定結果を「正解」と位置づける。その上で、2020年度（以下、「最新年度」）の各文書について、過年度文書の中から同じ内容のものだと考えられる文書を特定し、特定した文書に付された「正解」をもとに、当該文書の扱い（「廃棄」又は「保留」）を自動判定する。その後、自動判定の結果と、委員による実際の判定（＝「正解」）を比較し、自動判定の精度を評価する。

2.2　部門別の文書の仕分け

最新年度の個々の文書に対して過年度の全ての文書（約14万件）を参照することは効率的ではない。新原俊樹[13]は、行政機関の各課で保管されている文書が互いにどの程度類似しているか測定し、異なる地域であっても同じ業務を所掌する課の間では文書のタイトルが一致するものが多くなる傾向があることを明らかにしている。そこで本研究では、最新年度のある文書の自動判定を行う際に、当該文書を作成した課（又は、同じ業務を所掌する複数の課）の過年度文書に限定して参照することとし、熊本県の知事部局行政機構図[14]に基づき、本庁については各課単位で、広域本部と地域振興局については異なる地域で同じ業務を所掌する複数の課（以下、「部門」[15]）単位で、全ての文書を仕分けした[16][17]。仕分け後の各課・各部門の文書数を表１に示す。仕分けに当たり、熊本県庁処務規程[18]と熊本県広域本部処務規程[19]に基づき組織改編前後の連続性も考慮した。組織改編に伴い連続性が確認できなかった課については表１中に※印を付し、今回の調査対象から除外した。なお、表１中の「自動判定」列と「判定率」「再現率」「適合率」の各列については第４章で後述する。

2.3　コサイン類似度による文書の同一判定の考え方

次に、個々の文書の同一判定の考え方について説明する。本研究が採用したコサイン類似度による文書の同一判定の手法は次のとおりである。ある２つの文章があったとき、その中に含まれる単語の種類と数に応じたベクトルをそれぞれ定義する。この２つのベクトルの成す角がθであるとき、２つの文章のコサイン類似度は、その余弦値（$\cos\theta$）として得られる。ここ

表１　部門別の文書数と自動判定結果

No.	課名/部門名	過年度文書数	最新年度文書数	自動判定							判定率(%) (TP~TN)/全体	再現率(%) TP/(TP+FN)	適合率(%) TP/(TP+FP)
				閾値	TP	FP	FN	TN	UP	UN			
1	知事公室付	69	26	0.80	0	0	0	15	0	11	58	−	−
2	秘書グループ	131	115	0.75	1	2	0	52	1	59	48	100	33
3	広報グループ	413	131	1.00	2	6	1	46	13	63	42	67	25
4	くまモングループ	96	15	0.75	0	0	0	12	0	3	80	−	−
5	危機管理防災課	603	147	0.98	0	1	2	53	2	89	38	0	0
6	人事課	802	64	0.88	0	3	0	48	2	11	80	−	0
7	財政課	239	45	0.75	0	1	0	40	0	4	91	−	0
8	県政情報文書課	1268	258	1.00	1	4	0	80	7	166	33	100	20
9	総務厚生課	921	22	0.85	0	0	0	15	0	7	68	−	−
10	財産経営課	970	175	0.95	0	0	0	71	2	102	41	−	−
11	私学振興課	1242	384	0.98	7	12	1	96	10	258	30	88	37
12	市町村課	1418	466	0.75	32	62	6	216	2	148	68	84	34
13	消防保安課	741	241	0.83	1	4	1	137	4	94	59	50	20
14	消防学校	158	106	0.75	0	0	0	41	3	62	39	−	−
15	防災消防航空センター	88	51	0.75	1	0	0	25	0	25	51	100	100
16	税務課	457	223	0.85	11	12	2	64	15	119	40	85	48
17	自動車税事務所	204	67	0.83	0	0	1	23	0	43	36	0	−
18	企画課	753	299	0.85	3	5	5	121	10	155	45	38	38
19	東京事務所	526	74	0.75	3	1	4	50	0	16	78	43	75
20	地域振興課	576	266	0.90	4	9	4	94	7	148	42	50	31
21	文化企画・世界遺産推進課	344	68	1.00	0	0	0	10	4	54	15	−	−
22	博物館ネットワークセンター	66	35	0.75	0	0	0	12	0	23	34	−	−
23	交通政策課	165	75	0.85	1	1	1	20	7	45	31	50	50
24	統計調査課	828	159	0.95	2	8	18	64	15	52	58	10	20
25	球磨川流域復興局付	0	42										
26	健康福祉政策課	1374	375	0.83	14	20	7	183	17	134	60	67	41
27	福祉事務所	2442	496	0.78	8	43	2	332	5	106	78	80	16
28	保健所	7235	1799	0.80	103	275	42	965	36	378	77	71	27
29	福祉総合相談所	3028	524	0.75	11	19	4	387	4	99	80	73	37
30	保健環境科学研究所	479	101	0.93	0	3	0	60	0	38	62	−	0
31	健康危機管理課	881	240	0.75	15	23	6	140	6	50	77	71	39
32	食肉衛生検査所	436	112	0.93	5	8	3	46	23	27	55	63	38
33	高齢者支援課	1162	701	0.83	4	29	2	323	12	331	51	67	12
34	認知症対策・地域ケア推進課	321	246	0.90	2	5	0	55	14	170	25	100	29
35	社会福祉課	2289	480	0.78	14	18	6	315	21	106	74	70	44
36	子ども未来課	1076	717	0.75	11	11	13	321	14	347	50	46	50
37	子ども家庭福祉課	254	76	0.93	3	0	1	15	7	50	25	75	100
38	児童相談所	229	89	1.00	0	3	0	27	2	57	34	−	0
39	清水が丘学園	237	105	0.75	0	0	1	66	0	38	64	0	−
40	障がい者支援課	2079	553	0.75	12	31	18	298	15	179	65	40	28
41	精神保健福祉センター	523	171	0.83	1	6	6	143	2	13	91	14	14
42	こども総合療育センター	807	240	0.78	7	16	5	137	12	63	69	58	30
43	医療政策課	1581	448	0.75	5	37	12	260	10	124	70	29	12
44	国保・高齢者医療課	1057	139	0.95	5	0	0	47	13	74	37	100	100
45	健康づくり推進課	1370	633	0.75	11	15	22	231	17	337	44	33	42
46	薬務衛生課	1100	228	0.75	7	36	6	133	6	40	80	54	16
47	環境政策課	190	41	0.75	0	0	1	25	4	11	63	0	−
48	水俣病保健課	408	96	0.75	0	0	0	74	0	22	77	−	−
49	水俣病審査課	301	53	0.75	0	0	0	40	0	13	75	−	−
50	環境立県推進課	397	161	0.88	1	2	0	88	3	67	57	100	33
51	環境センター	306	74	0.75	0	4	0	47	0	23	69	−	0
52	環境保全課	779	163	0.98	5	6	7	79	10	56	60	42	45
53	自然保護課	452	123	0.98	4	1	0	30	11	77	28	100	80
54	循環社会推進課	671	221	1.00	1	0	3	67	9	141	32	25	100
55	くらしの安全推進課	476	126	0.75	2	8	1	73	2	40	67	67	20
56	消費生活課	324	91	0.83	7	3	3	39	3	36	57	70	70
57	男女参画・協働推進課	612	156	0.90	0	3	1	106	0	46	71	0	0
58	人権同和政策課	400	102	0.78	9	11	5	38	3	36	62	64	45
59	商工政策課	252	134	0.78	2	5	1	36	8	82	33	67	29
60	大阪事務所	217	58	0.80	1	2	0	41	1	13	76	100	33
61	福岡事務所	43	10	0.75	0	0	0	7	0	3	70	−	−
62	商工振興金融課	508	122	0.83	3	2	1	56	9	51	51	75	60
63	労働雇用創生課	1037	480	0.98	8	6	8	70	69	319	19	50	57
64	高等技術専門校	495	123	0.80	0	1	0	100	0	22	82	−	0
65	技術短期大学校	554	194	0.88	2	1	0	69	0	122	37	100	67
66	産業支援課	402	170	1.00	5	7	2	55	13	88	41	71	42
67	産業技術センター	692	189	0.78	0	1	0	148	1	39	79	−	0
68	エネルギー政策課	178	57	0.85	2	1	1	21	2	30	44	67	67
69	企業立地課	468	161	0.95	5	1	0	36	3	116	26	100	83
70	観光交流政策課	0	82										
71	観光企画課	0	142										
72	観光振興課	194	8										
73	販路拡大ビジネス課	0	22										
74	農林水産政策課	1021	155	0.93	6	2	0	62	2	83	45	100	75

表1　部門別の文書数と自動判定結果（続き）

No.	課名/部門名	過年度文書数	最新年度文書数	自動判定							判定率(%) (TP～TN)/全体	再現率(%) TP/(TP+FN)	適合率(%) TP/(TP+FP)
				閾値	TP	FP	FN	TN	UP	UN			
75	団体支援課	1495	373	0.95	23	21	7	143	20	159	52	77	52
76	流通アグリビジネス課	236	71	0.93	0	1	1	20	2	47	31	0	0
77	農業技術課	985	159	0.83	0	3	4	68	7	77	47	0	0
78	農業研究センター	1136	475	0.83	3	6	6	299	7	154	66	33	33
79	病害虫防除所	23	6	0.75	0	2	0	4	0	0	100	−	0
80	農産園芸課	750	232	1.00	7	10	2	77	11	125	41	78	41
81	畜産課	1410	378	1.00	10	7	3	135	24	199	41	77	59
82	家畜保健衛生所	2324	612	0.83	20	28	20	409	12	123	78	50	42
83	農地・担い手支援課	693	281	0.80	8	19	4	115	11	124	52	67	30
84	農業大学校	852	0										
85	農村計画課	525	145	0.85	2	6	1	66	0	70	52	67	25
86	農地整備課	288	128	1.00	0	0	0	32	3	93	25	−	−
87	大切畑ダム復興事務所	1	4	0.75	0	0	0	1	0	3	25	−	−
88	むらづくり課	397	261	1.00	5	1	2	21	23	209	11	71	83
89	技術管理課	248	114	0.80	0	1	8	51	2	52	53	0	0
90	森林整備課	794	246	0.75	5	10	9	144	10	68	68	36	33
91	林業研究・研修センター	298	84	1.00	1	1	2	50	0	30	64	33	50
92	林業振興課	732	380	0.93	14	7	11	116	21	211	39	56	67
93	森林保全課	485	113	0.75	9	8	3	61	1	31	72	75	53
94	水産振興課	1255	133	0.85	2	5	2	61	5	58	53	50	29
95	水産研究センター	218	62	0.90	5	1	7	9	11	29	35	42	83
96	漁港漁場整備課	726	72	0.80	1	1	0	36	2	32	53	100	50
97	漁業取締事務所	202	60	0.78	1	0	0	31	1	27	53	100	100
98	監理課	720	337	0.90	4	2	8	133	18	172	44	33	67
99	用地対策課	163	25	0.75	3	1	1	13	0	7	72	75	75
100	土木技術管理課	156	59	0.75	0	0	2	34	6	17	61	0	−
101	道路整備課	744	144	0.80	1	4	1	75	1	62	56	50	20
102	道路保全課	232	132	0.75	0	0	3	32	16	81	27	0	−
103	都市計画課	423	147	1.00	3	0	3	34	10	97	27	50	100
104	下水環境課	322	155	0.83	3	3	1	64	6	78	46	75	50
105	河川課	1022	685	0.75	9	18	16	169	119	354	31	36	33
106	ダム管理所	105	31	0.75	1	0	1	21	1	7	74	50	100
107	港湾課	197	35	0.90	1	0	2	8	1	23	31	33	100
108	港管理事務所	638	188	0.75	2	4	0	109	3	70	61	100	33
109	天草空港管理事務所	260	63	0.75	0	3	0	33	0	27	57	−	0
110	砂防課	332	34	0.75	2	0	0	19	2	11	62	100	100
111	建築課	1095	363	0.78	5	8	6	242	1	101	72	45	38
112	営繕課	2171	414	0.88	0	0	1	96	0	317	23	0	−
113	住宅課	1585	489	0.75	1	3	6	249	24	206	53	14	25
114	総務（総務）	1490	143	0.75	4	1	5	26	18	89	25	44	80
115	総務・振興	6327	1870	0.75	65	67	23	1301	35	379	78	74	49
116	課税・収税	12964	2439	0.75	30	411	29	1478	17	474	80	51	7
117	総務・企画・福祉	4605	1520	0.85	32	101	40	591	93	663	50	44	24
118	衛生環境	217	32	0.75	0	0	0	11	2	19	34	−	−
119	保健予防	366	52	0.93	0	0	0	24	0	28	46	−	−
120	試験検査	8	2	0.75	0	0	0	2	0	0	100	−	−
121	総務（農林）	396	107	0.98	0	0	0	56	0	51	52	−	−
122	農業普及・振興	5335	1452	0.78	167	270	83	613	46	273	78	67	38
123	農地整備	2902	382	0.83	3	17	7	186	8	161	56	30	15
124	土地改良	912	54	0.75	0	2	3	20	1	28	46	0	0
125	林務	4196	1255	0.78	120	221	26	525	36	327	71	82	35
126	森林保全	457	294	0.75	7	3	10	121	16	137	48	41	70
127	山地災害対策	52	23	0.95	1	0	0	2	3	17	13	100	100
128	水産	311	72	1.00	3	2	3	24	13	27	44	50	60
129	漁港	126	4	0.75	0	0	0	2	2	0	50	−	−
130	総務（土木）	1715	76	0.75	1	0	0	61	1	13	82	100	100
131	総務出納	129	46	0.75	0	0	0	31	0	15	67	−	−
132	技術管理	1015	76	0.75	2	1	1	42	2	28	61	67	67
133	景観建築	705	76	0.75	0	0	4	54	1	17	76	0	−
134	用地	841	248	0.75	5	6	5	175	1	56	77	50	45
135	工務	3137	531	0.75	0	2	0	95	1	433	18	−	0
136	維持管理	2289	768	0.75	1	4	9	433	4	317	58	10	20
137	会計課	1424	349	0.78	2	1	4	261	11	70	77	33	67
138	管理調達課	1095	277	0.90	4	0	2	82	3	186	32	67	100
139	川辺川ダム総合対策課（※）	124	0										
140	総務事務センター（※）	700	494										
141	情報政策課（※）	287	7										
142	情報企画課（※）	390	85										
143	国際課（※）	267	0										
144	くまもと県民交流館（※）	26	0										
145	熊本駅周辺整備事務所（※）	29	0										
146	（不明）	1711	91										
	合計	137201	36184		968	2089	633	16792	1188	13413	58	60	32

では、文書のタイトルを１つの文章としてベクトルを定義する。通常、文書のタイトルは数十字程度の短い文章であり、文書の本文の文字数と比較しても圧倒的に少ないが、その内容を表現するのに不可欠な単語が精選されて決められている。また、管理簿を通じて全ての文書のタイトルを容易に入手できるほか、比較する文書の媒体（紙又はファイル）やファイル形式の違いを考慮する必要がないという利点がある。２つの文書のタイトルを比較してコサイン類似度を算出する一連の手順を図１に示す。

図１　コサイン類似度の算出手順

コサイン類似度の算出に先立ち、各文書のタイトルに対応したベクトルを定義するため、部門別に、全ての文書タイトルに含まれる単語とそれらの出現回数を集計した。その上で、集計した単語と同数の次元を有するベクトル空間上に、各タイトルに対応するベクトルを配置した（図１の手順１）。配置したベクトルの各成分の値については、各文書のタイトルに含まれる各単語の出現頻度を示す指標（TF：Team Frequency）と、全文書を通じた各単語の希少性を示す指標（IDF: Inverse Document Frequency）[20]の積により決定した（図１の手順１）。これらのベクトルの成す角の余弦値（cos θ）が、２つの文書間のコサイン類似度となる（図１の手順２）。２つのタイトルに共通する単語が全く含まれない場合、対応する２つのベクトルの内積は０となり、コサイン類似度も０になる。一方、２つの文書のタイトルが全く同じである場合、対応する２つのベクトルは全く同じ値となり、コサイン類似度は１になる。すなわち、コサイン類似度は０から１の値をとり、１に近いほど２つの文書のタイトルには共通の単語が多く含まれていることになる。

ここで算出したコサイン類似度は、文書のタイトル同士の類似性を示したものであり、コサイン類似度の高さが必ずしも文書の内容の一致を保証するものではない。しかし、新原俊樹[13]の調査では、コサイン類似度が高い文書の組合せほど、その内容が一致する組合せが増える傾向にあり、例えば、コサイン類似度が0.85となる文書の組合せのうち、67％の組合せで内容が一致することが報告されている。文書が同一であると判定するコサイン類似度の閾値をどの値に設定するのが妥当か、3.3節であらためて検証する。

なお、本研究の一連の分析には、機械学習の分野で実績のあるプログラミング言語の Python を使用し、形態素解析器には MeCab を使用した[21][22]。また、TF-IDF とコサイン類似度の計算には、Python を使って機械学習を行うためのプログラム群である scikit-learn[24]を用いた。

2.4　文書タイトルの補正

文書タイトルのベクトル化に先立ち、タイトル内の表現について補正を行った。これは、文書の内容に直接関係しない語を含んだままベクトル化することで、ベクトル間のコサイン類似度が不必要に低下してしまうことを防ぐ意図がある。具体的には、全ての英数字を半角に統一したほか、冗長的な表現（例えば、「～に関する文書」「～関係」「～関連」「～総記」「～綴」など）の省略、元号（平成・令和）や通番（No.１、第２号など）の削除を行った。また、文書のタイトルに続けてサブタイトルが付与されている文書については、両者を連結して１つのタイトルとした。

2.5　過年度文書の第一次選別結果の補正

過年度文書の中には、同じ名称の文書が複数年に跨って作成されるものも多いが、年度によって第一次選別の結果が異なる事例も少なくない。こうした齟齬が生じるには、以下の原因がある。

（1）初期の年度に「保留」と判定して第二次選別（現物確認）を行ったが、現物確認の結果、「廃棄」と判定され、以降の年度の同様の文書が第一次選別で「廃棄」と判定されるようになった。

（2）初期の年度に「廃棄」と判定されていたが、その後、当該文書の重要性が認識され、以降の年度の同様の文書が第一次選別で「保留」と判定されるようになった。

（3）ある年度に「保留」と判定すべき文書が見落とされ、「廃棄」と判定されていた。

いずれにしても、「正解」となるべき第一次選別の結果に齟齬がある状態では、最新年度の文書の自動判定の参考にすることができない。そこで、本来、永久保存とすべき文書を漏れなく判定する観点から、第一次選別の結果が異なる文書については、すべて「保留」に統一することとした[25)]。

具体的な補正手順は、部門別に、過年度文書を「保留」のグループと「廃棄」のグループに分けた上で、各文書と相手グループの全文書との間でそれぞれコサイン類似度を算出する。この値が一定の閾値以上の場合、2つの文書は同一であると判定し、「廃棄」グループに置かれている方の文書を「保留」グループに移すとともに、第一次選別結果も「廃棄」から「保留」に補正する。この操作を何度か繰り返すと、「保留」グループに置かれた文書と同一と判定された文書の選別結果は、全て「保留」に統一され、「正解」に含まれている判定結果の齟齬が解消される。

2.6　第一次選別の自動判定

最後に、2.5節で得た補正済みの「正解」を参照して、最新年度文書の扱い（「保留」又は「廃棄」）を自動で判定する。具体的には、最新年度の個々の文書について、当該部門に仕分けされた過年度の全文書との間でコサイン類似度を算出する。算出したコサイン類似度が閾値以上となった文書に付された判定結果（＝補正済みの「正解」）を参照し、当該文書の判定結果とする。このとき、もし、コサイン類似度が閾値以上となる文書が複数あり、それらの判定結果が異なる場合は、「保留」の結果を優先する。なお、過年度文書の中にコサイン類似度が閾値以上となるものが存在しなかった場合、自動判定の結果は「不明」とした。

3　市町村課を事例とした自動判定の結果と精度評価

事例として、市町村課における最新年度の文書の自動判定結果を確認し、その精度について評価した。市町村課は本庁の総務部のみに置かれた課であり、市町村の行財政や地方創生に関する業務を所掌している[18)]。同課で作成される文書は毎年開催される会議に関するものが多いため、最新年度文書の自動判定を行うに当たって、比較的、過年度文書の判定結果が参照しやすい。なお、自動判定に当たり、文書が同一であるとみなすコサイン類似度の閾値は0.75とした（この値の妥当性は3.3節で検証する）。

3.1　自動判定の結果

市町村課で過年度（2016～2019年度）に保存期間満了を迎えた文書は1,418件あり、そのうち、第一次選別で委員に「保留」と判定されていた文書は237件（全体の16.7%）であった。2.5節の手順に基づき、この結果を補正したところ、「保留」の文書数は437件（同30.8%）に増えた。この補正済みの「正解」を用いて、同課の最新年度（2020年度）の文書466件について、2.6節の手順で自動判定を行った。自動判定の結果と、委員による実際の判定結果を表2に整理した[26)]。

表2の各欄に整理した値から、自動判定の精度を測るための各種の指標を算出することができる。まず、表2中の各欄の意味を整理する。TP（True Positiveの略）は、委員が「保留」と判定した文書を正しく「保留」と予測できた文書数を示し、TN（True Negativeの略）は、委員が「廃棄」と判定した文書を正しく「廃棄」と予測した文書数を示す。一方、FN（False Negativeの略）は、委員が「保留」と判

表２　市町村課の最新年度文書の自動判定結果（同一判定閾値０.75の場合）

		自動判定結果		
		「保留」	「廃棄」	「不明」
委員判定結果	「保留」	TP 32	FN 6	UP 2
	「廃棄」	FP 62	TN 216	UN 148

定した文書を誤って「廃棄」と予測した文書数であり、FP（False Positive の略）は、委員「廃棄」と判定した文書を誤って「保留」と予測した文書数である。なお、コサイン類似度が閾値以上となる過年度文書が存在せず「保留」又は「廃棄」の予測ができなかった文書のうち、委員が「保留」としていた文書数を「UP（Unknown Positive の略）」欄に、「廃棄」としていた文書数を「UN（Unknown Negative の略）」欄に記載した。

3.2　自動判定精度の評価

自動判定の精度を評価するため、表２の各欄の値から各種の指標を算出した。まず、（TP+FP+FN+TN）/（TP+FP+FN+TN+UP+UN）の式から求めた「判定率」は58％であった。これは、最新年度の全文書466件のうち、自動で「保留」又は「廃棄」の予測が出来た文書（316件）の比率を示している。この比率が高いほど、自動判定できる文書の割合が増え、人間が第一次選別を行う文書の数が減るので、作業効率は高まる。しかし、同一判定の閾値を下げることで判定率を高めることはできるが、必ずしも正確な判定には繋がらないことに留意が必要である。

次に、（TP+TN）/（TP+FP+FN+TN）の式から求められる「正解率（Accuracy）」は78.5％であった。これは、自動判定を行うことができた全文書316件（分母）に対して、実際の委員の判定を正しく予測できた割合である。一般的には、この値が高いほど自動判定の精度が良いとされる。しかし、316件のうち、実際に委員から「廃棄」と判定された文書は278件（88.0％）に上るため、仮に全ての文書を「廃棄」と予測しても正解率は90％近くになってしまう。今回の事例においては、正解率を判定精度の評価指標に用いることは適切ではない。

また、TP/（TP+FN）の式から得られる「再現率（Recall）」は84.2％であった。これは、委員が「保留」と判定した文書（分母）に対して、どれほど取りこぼしなく自動判定で「保留」と判定できたかを示す指標である。再現率は真陽性率（TPR: True Positive Rate）とも呼ばれる。この第一次選別においては、本来「保留」とすべき文書を確実に捕捉することが求められるため、再現率には高い数値が求められる。一方で、過年度の同じ名前の文書は「廃棄」とされていながら、最新年度に委員が初めて「保留」と判定した文書も少なからず存在する。こうした事例もあるため、過年度の結果に基づき自動判定する本手法では再現率を100％に近づけることは難しい。

さらに、TP/（TP+FP）の式から算出される「適合率（Precision）」は34.0％であった。これは、自動判定で「保留」と予測した全文書（分母）に対して、実際に委員が「保留」と判定した文書の割合を示したものである。適合率が低いことは、真に「保留」とすべき文書の数に対して、自動判定で「保留」と予測した文書の数が多すぎるということであり、第二次選別（現物確認）に回る文書の数が必要以上に多くなっており、評価選別の作業全体として非効率な状態であることを意味している。一方で、本来、「保留」とすべきものが見逃されて「廃棄」とされている文書を捕捉している側面もあり、適合率が低いことが一概に不適切な状態とも言えない。

続いて、FP/（FP+TN）の式から得られる偽陽性率（FPR: False Positive Rate）は22.3％であった。これは、委員が「廃棄」と判定した文書（分母）に対して、自動判定で誤って「保留」と予測した割合を示す。第一次選別で本来「廃棄」と判定されるべき文書を誤って「保留」と予測することで、第二次選別における現物確認の手間をいたずらに増やすことになるため、この指標はより低いことが望ましい。

3.3　コサイン類似度の閾値の調整

3.2節で求めた各指標の値は、文書が同一であると判定するコサイン類似度の閾値を0.75とした場合に得

られたものである。この閾値を変えることで表２の各欄に入る文書の数が変わり、それぞれの指標の値も変化する。仮に、閾値を著しく上げると、文書の同一判定の基準が厳しくなり、過年度の同じ文書の判定結果をより正確に参照できるようになる一方、自動判定可能な文書の数が減って判定率の低下を招く。逆に、閾値を大幅に下げると、人間が命名する文書ファイル名の表現の揺らぎを許容することができ、判定率は高くなるが、別の内容の文書を同一と判定してしまう危険性が高まり、判定精度の低下に繋がる。各指標がバランス良く望ましい値に近づくように、最適な閾値を設定する必要がある。

そこで、本研究では、再現率と偽陽性率に注目し、偽陽性率を低く抑えつつ、可能な限り再現率が高くなるように閾値を選定する。図２は、縦軸に再現率、横軸に偽陽性率を取り、市町村課の文書について閾値を変化させながら得られた両指標を描画したものである[27)]。理想的には、偽陽性率が限りなく０％に近く、再現率が限りなく100％に近づいている状態、すなわち、図２中の左上隅（★印の位置）ほど判定精度が高いことを意味する。市町村課の事例では、図中の各点の位置から判断し、コサイン類似度の閾値が0.75の点が図の左上隅に最も近く（図２中に点線で表示）、妥当な閾値になっている。ただし、この結果は、他の部門の文書にそのまま適用することはできない。部門別に最適な閾値を探索する必要がある。

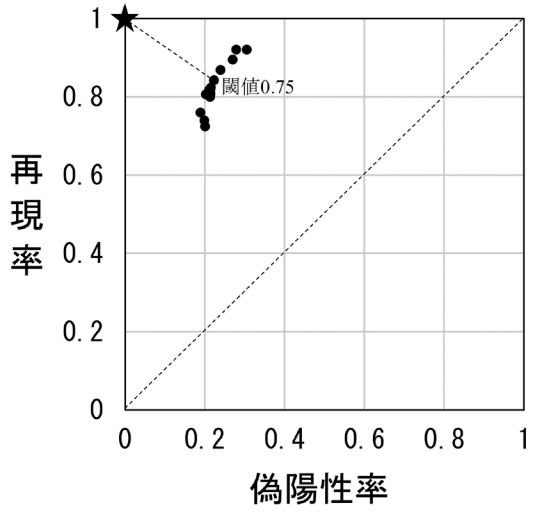

図２　それぞれの閾値に対する再現率と偽陽性率の分布

4　全部門の結果と判定精度の向上に向けた取組

続いて、市町村課以外の各部門についても、3.3節の手法で最適な閾値を設定した上で自動判定を行った。自動判定の結果として、TP, FP, FN, TN, UP, UN の各欄に振り分けられた文書の数と、これらの数から算出した「判定率」「再現率」「適合率」の各指標を表１に整理した。なお、自動判定の結果によっては算出できなかった指標もあった。その場合、該当する指標には「－」を付した。また、全部門を通じて TP, FP, FN, TN, UP, UN に分類された文書を集計し、各指標を算出したところ、全体の判定率は58％、再現率は60％、適合率は32％となった。いずれも高くなることが理想であるが、それぞれの指標はトレードオフの関係にあり、全ての指標を同時に改善することは容易ではない。判定精度の向上に向けて、今後、検討すべき課題について述べる。

4.1　その他の機械学習モデルによる自動判定手法の検討

本研究では、コサイン類似度に基づく同一判定手法を用いたが、機械学習を用いて「廃棄」や「保留」など二つの値に分類する手法としては、このほかにも「決定木」や「ランダムフォレスト」[28)]など様々な手法がある。これらの手法による判定精度を比較し、最適な手法と閾値等のパラメーターを決定することで、判定精度を高めていくことができる。

4.2　ファイル命名則の改善

第一次選別は文書のタイトルに基づいて行うため、文書の作成時にその内容を正確に表現したタイトルを付与することが判定精度の向上に直結する。しかし、現状では、同じ内容でありながら異なるタイトルが付与されていたりタイトル名に誤植が含まれていたりするケースや、逆に、内容が異なるにもかかわらず、類似したタイトルが付与されているために内容が違うことが把握できないケースが多々ある。文書に名称を付与する立場の職員が、同じ案件の文書に同じタイトル

を付与し、また、サブタイトルを活用して各文書の詳細を的確に表現する能力を高めるための支援が不可欠である。本研究の成果の一つとして、部門別の文書一覧を作成したところであり、各部門の職員が一覧を活用して文書のタイトルを付与することにより、過年度文書と同じ内容の文書に同じタイトルが付されるようになれば、判定率も大きく向上するだろう。

4.3　過年度における第二次選別結果の反映

2.5節「過年度文書の第一次選別結果の補正」において、同じ名称ながら第一次選別の結果が異なる文書は全て「保留」に統一した。しかし、この運用のままでは、過年度文書が増えるにつれて「保留」とされる文書が全体に占める割合が徐々に増えていく。このため、補正済みの「正解」に基づいて自動判定を行うと、FP（委員が「廃棄」とした文書のうち、自動判定で「保留」としたもの）に分類される文書の数が増え、第二次選別で現物確認をする手間が増えていく。現行の自動判別の手法では、過去のある年度の第一次選別の際に一度だけ「保留」とされた文書があった場合、その後の現物確認で「廃棄」と判断されていたとしても、最新年度の文書の自動判定で当該文書と同一と判定された文書は全て「保留」と判断され、再び現物確認が求められる。

この問題を回避するためには、過年度の第二次選別結果を第一次選別の過程に反映していくことが必要である。具体的には、第一次選別で「保留」と判定されて第二次選別に回された文書について、現物確認後に明らかになった「永久保存」又は「廃棄」の結果を補正後の第一次選別結果に還元する。これにより、次年度以降の自動判定の際に「永久保存」の文書と同一であると判定されたものは、第二次選別を経ずに「永久保存」と判定できるようになる。

5　まとめと今後の課題

熊本県の知事部局において、2016～2020年度の間に保存期間満了を迎えた文書を対象とし、過年度文書（2016～2019年度）の第一次選別結果に基づき、コサイン類似度に基づく文書の同一判定手法を用いて、最新年度文書（2020年度）の第一次選別の自動判定を行った。委員による実際の判定結果と比較して自動判定の精度を評価したところ、判定率は全体の58％、再現率は60％、適合率は32％となった。各指標はトレードオフの関係にあり、全ての指標を共に改善することは容易ではない。判定精度の改善に寄与し得るものとして、以下の３つの研究課題がある。

（１）機械学習を用いた２値分類の手法としては、コサイン類似度に基づく同一判定以外にも、決定木やランダムフォレストなど様々な学習モデルがある。これらのモデルによる判定精度を比較し、最適な学習モデルとパラメーターを決定することで、判定精度を高めることができる。

（２）現状、４割近くの文書は「過年度に同一の文書がない」とされ、自動判定を行えていない。過年度の文書名にも配慮し、ファイルの内容に合わせて適切な名称を付与するしくみが必要である。これにより、判定率が高まり、目視で第一次選別を行う文書の数を減らすことができる。

（３）過年度の第二次選別の結果、「永久保存」と判定された文書の情報を第一次選別の過程に反映していくことで、次年度以降の自動判定の際に、第二次選別を経ずに「永久保存」と判定する文書を増やすことができる。自動判定の精度を高めるには、特に、第二次選別結果の反映が不可欠であり、今後の重要な研究課題である。

謝辞

本研究は、2022年度記録管理学会プロジェクト研究「機械学習を用いた行政文書の第一次選別機能の開発と性能評価」の成果をまとめたものである。なお、本研究の分析対象である熊本県の第一次選別結果の利用に当たり、同県の許可を得た。この内容の主要な部分は、記録管理学会2023年研究大会（於 お茶の水女子大学、2023年５月27日）において発表している。この発表に当たり、JSPS 科研費23H03692「熊本県の事例研究に基づく地方公共団体の模範となる文書管理モデルの確立」の助成を受けた。本研究に助成いただいた記録管理学会にお礼申し上げます。

注

1）電子政府の総合窓口イーガブ「公文書等の管理に関する法律」，https://elaws.e-gov.go.jp/search/elawsSearch/elaws_search/lsg0500/detail?lawId=421AC0000000066，（参照2023-08-14）.

2）地方公共団体公文書管理条例研究会．“地方公共団体における公文書管理とは”．こんなときどうする？自治体の公文書管理～実際にあった自治体からの質問36．第一法規株式会社，2019，pp. 2-17.

3）地方自治研究機構「公文書管理に関する条例」，http://www.rilg.or.jp/htdocs/img/reiki/019_officialdocumentmanagement.htm，（参照2023-08-14）.

4）ただし、他の道府県は規則・規程・要綱等の形で制定している。

5）宮間純一．“公文書管理法後の自治体の文書管理”．公文書管理法時代の自治体と文書管理．勉誠出版，2022，pp. 3-22.

6）熊本県「熊本県行政文書等の管理に関する条例」，https://www1.g-reiki.net/kumamoto/act/frame/frame110000040.htm，（参照2023-08-14）.

7）楠本誠二．熊本県における行政文書管理制度．アーカイブズ．2014，vol. 52，pp. 66-69.
https://www.archives.go.jp/publication/archives/wp-content/uploads/2015/03/acv_52_p66.pdf

8）三輪宗弘．“何を残すべきなのか－熊本県公文書への私のチャレンジと日本への提言－”．アーカイブズとアーキビスト－記録を守り伝える担い手たち－．大阪大学出版会，2021，pp. 93-131.

9）新原俊樹．地方公共団体の出先機関における標準文書保存期間基準（保存期間表）の作成．レコード・マネジメント．2021，vol. 80，pp. 35-46.
https://doi.org/10.20704/rmsj.80.0_35

10）緒方克治，牧尾亘．キーワードから考察する公文書の選別と廃棄－熊本県の第一次選別手法の試み．記録管理学会2022年研究大会予稿集．2022，pp. 11-15.

11）永村美奈．熊本県における特定歴史公文書の評価選別の分析．レコード・マネジメント．2015，vol. 69，pp. 87-103.
https://doi.org/10.20704/rmsj.69.0_87

12）永井リサ．熊本県における農林業関連行政文書の管理保存について－「農林業センサス」を中心に－．記録管理学会2022年研究大会予稿集．2022，pp. 39-43.

13）新原俊樹．行政文書の分類・整理に係る支援機能の提案．レコード・マネジメント．2018，vol. 75，pp. 48-59.
https://doi.org/10.20704/rmsj.75.0_48

14）熊本県「熊本県知事部局行政機構図」，https://www.pref.kumamoto.jp/uploaded/life/175360_421010_misc.pdf，（参照2023-08-14）.

15）例えば、農業普及・振興課は、県北・県央・県南・天草の各広域本部と、玉名・鹿本・阿蘇・宇城・上益城・芦北・球磨の各地域振興局内にそれぞれ置かれている。これらの課には、互いに類似する文書が多く保管されており[13]、互いに過年度文書の第一次選別結果が参考になることから、これらの課の文書を一つの「農業普及・振興」部門の文書としてまとめた。

16）出先機関によって組織構成が異なるため、各課の文書を部門別に仕分けする際には、複数の課の文書をまとめて１つの課の文書として集計した事例（例えば、総務課と振興課の文書をまとめて「総務・振興」部門の文書とみなした）がある。

17）文書を部門別に仕分けするに当たり、文書の所属組織名の記載の整合が取れていない事例が散見された。特に、地域振興局内の各課の名称を略称で記載する場合に、省略方法の整合が取れていないケースや、地方振興局としての文書と広域本部としての文書の扱いが混同されているケース、また、いずれの課でもなく所属機関の長の所属とされるケースもあった。これらの文書については目視で確認して可能な限り仕分けを行ったが、どの課に所属するか判別できなかったものについては「不明」とした。文書の所属課については、文書を行政文書管理システムに登録する際に、事前に作成された選択肢の中から選ぶ方式にするなど、文書の登録者によって齟齬が生じないようにする工夫が必要である。

18）熊本県「熊本県庁処務規程」，https://www1.g-reiki.net/kumamoto/act/frame/frame110000183.htm，（参照2023-08-14）.

19）熊本県「熊本県広域本部処務規程」, https://www1.g-reiki.net/kumamoto/act/frame/frame110010146.htm, （参照2023-08-14）.

20）IDF は、全文書数 N を各単語の出現回数 n で割った値の自然対数、すなわち、loge（N/n）によって求めた。出現回数 n が少ない単語ほど、IDF は大きくなる。

21）「MeCab: Yet Another Part-of-Speech and Morphological Analyzer」, http://taku910.github.io/mecab/, （参照2023-08-14）.

22）具体的には、MeCab を Python で利用するためのプログラム群である mecab-python3[23] を使用した。

23）Python Package Index「mecab-python3 1.0.6」, https://pypi.org/project/mecab-python3/, （参照2023-08-14）.

24）Pedregosa et al. 'Scikit-learn: Machine learning in Python,' *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830.

25）技術的には、「廃棄」又は「保留」のいずれかの分類ではなく、過年度に同一と判定された文書が複数ある場合に、その中で「保留」と判定された文書の割合（以下、「保留率」）として得ることも可能である。しかし、行政実務の現場では、仮に連続値として保留率が得られても、それが 0 %でない限り現物確認をせざるを得ないのが実情である。

26）表 2 は、機械学習の分野では混同行列（Confusion Matrix）と呼ばれる。

27）図 2 は ROC 曲線（Receiver Operatorating Characteristic curve：受信者動作特性曲線）とも呼ばれ、自動判定の精度を評価する際によく利用される。

28）決定木は、ツリー構造を用いて判定する際の最適な判定要因を決める手法である。ランダムフォレストは、異なる複数の決定木で判定を行い、それぞれの判定結果の多数決をとる分析手法である。