教育社会学・教育学・社会学に着目した テキストマイニングによる学問分野間関係の分析

増田 勝也† 堤 孝晃‡ 齋藤 崇徳‡
†東京大学 知の構造化センター ‡東京大学大学院教育学研究科

masuda@cks.u-tokyo.ac.jp, takaaki.223@gmail.com,
bachianasbrasileirasbyhvl@gmail.com

1 はじめに

本研究では、学問分野間の位置関係を学会誌論文を対象としたテキストマイニングにより明らかにする。学問分野間の位置関係に関する研究は、社会科学系の学問領域において、分野の専門分化、学際化の様相を明らかにする目的で行われている。例えば本研究で対象とする「教育社会学」は「教育学」「社会学」という2学問分野を親学問として持つため、それらの親学問との関係に関する研究が多く行われている [1, 3]。しかしながら、それらの研究は一般的にある論者によるレビューという形を取ることが多く、十分な実証的検討がなされているとは言いがたい。そこで本研究では、学問分野間の関係をテキストマイニングを用いて実証的に分析することを目的とする。マイニングの対象は各学問分野を代表する論文誌とするが、多くの論文には機械で処理可能なデジタルテキストが存在しないため、論文画像からOCRシステムを用いてテキストを抽出し分析に利用する。分析では、各論文誌を用語、係り受け関係により特徴付け、頻出用語・係り受け関係の頻度分析、およびそれらの頻度を利用した文書集合間の類似度により、対象とする「教育社会学」「教育学」「社会学」の学問分野間の位置関係を明らかにする。

2 データと分析手法

2.1 分析対象とテキストデータの作成方法

本研究では日本教育学会・日本教育社会学会・日本社会学会それぞれの機関誌である『教育学研究』『教育社会学研究』『社会学評論』の3つの雑誌を対象に、テキストマイニング分析を行う[1]。各分野の研究内容を取り出すために、書評などは除外し、特集論文・投稿論文のみを対象とする。また対象年代を雑誌間で統一し1950年から2009年の60年分とする。各雑誌の分析対象論文数を表1に示す。分析のためのテキストデータは、東京大学知の構造化センターの「岩波書店『思想』の構造化プロジェクト」で開発された手法 [2] を用いて作成した。手法の概要は以下のとおりである。まず対象となる論文画像に対しOCRシステムによる認識を行った。その結果認識されたブロックに対し、機械学習手法を用いて属性分類を行った。対象とする属性は「論文タイトル」「章・節タイトル」「著者」「柱(ヘッダ、フッタ等)」「ページ番号」「本文」「注」「参考文献」「英文要旨」の9種類である。人手により全体のおよそ10%の353論文について正解データを作成し、学習に利用した。今回の分析においては、その中で「柱」および「ページ番号」を除外した、残りのすべての部分を分析対象とした[2]。なおOCRによる文字認識は完全ではなく、今回の処理における文字単位での精度をサンプリングデータに対し求めたところ約99.3%であった。とくに1950年代の論文においては元画像が不鮮明ということもあり、論文によっては約97%まで低下する。

表 1: 論文誌データ

	〈教育〉	〈教社〉	〈社会〉
対象論文数	1,414	821	1,436
総用語数	1,972,409	1,057,558	1,933,685
総係り受け関係数	803,129	468,602	776,190

[1]以降、簡便のため『教育学研究』を〈教育〉、『教育社会学研究』を〈教社〉、『社会学評論』を〈社会〉と呼ぶ。

[2]本来は本文のみを対象としたいが、現在の手法では、とくに参考文献や注の形式的な自動分類は困難で、多くが本文として分類されてしまう。そのため、分類精度によって分析対象の雑誌間、年代間で差異が生じることを避けるため、高い精度で分類できる「柱」および「ページ番号」のみを除外した。

表 2: 雑誌間類似度

	DF	TF
〈教育×教社〉	0.930	0.890
〈教社×社会〉	0.941	0.766
〈教育×社会〉	0.921	0.702

2.2 分析方法

前節の方法により作成したテキストデータを用いて、本論文では「用語」「係り受け関係」の頻度分析、「コサイン類似度」による分析を行う。「用語」は、対象テキストに対して形態素解析器 MeCab[3]を用いて形態素解析を行い、その結果における名詞の連続とする。今回はそのうち頻度上位の用語 18,000 語を抽出し、そのうち人手により分析に有用であると判定した 14,351 語を分析に使用する[4]。また「係り受け関係」は、対象テキストから係り受け解析器 CaboCha[5]により分節間の係り受け関係を求め、その係り元・係り先分節中の用語に対する全「係り元用語ー係り先用語」ペアを「係り受け関係」として抽出した。なお、今回は係り受け関係の方向は考慮しないものとする。各雑誌における用語および係り受け関係の総出現頻度を表 1 に示す。

頻度分析に用いる指標には用語・係り受け関係の TF、DF の両方を使用し、それぞれの頻度指標により異なる観点から分析を行う。また、雑誌間類似度の尺度には論文集合に対する「コサイン類似度」を利用する。すなわち、対象とする雑誌・年代に該当する論文集合に対し、その中での用語または係り受け関係の頻度を要素とするベクトルをそれぞれ作成し、そのベクトル間のコサイン値として類似度を算出する。

3 学問分野間関係の分析

3.1 雑誌間類似度による分析

用語の頻度 (TF、DF) を用いた雑誌間のコサイン類似度を表 2 に示す。TF、DF いずれの場合も〈教育×社会〉が最も類似度が低い。これは教育学と社会学という教育社会学の 2 つの親学問同士は「距離」が遠く、親子関係では「距離」が近いことを意味しており、妥当な結果であるといえる。残りの二つの組合せについては、TF では〈教育×教社〉の類似度が最も高いのに対し、DF では〈教社×社会〉が最も類似度が高いという違いがみられる。

3.2 頻出用語による分析

具体的な様相を把握するために、各雑誌における頻出語上位 20 語を表 3 に示す。DF では、3 誌に重複する用語が上位 20 のうちの半数以上の 11 を占め、具体的には、「問題」「意味」「重要」といった社会科学として共通に用いることの多い基礎的な用語であることがわかる。2 誌重複は、〈教育×社会〉が「理解」の 1 用語のみであるのにたいし、〈教育×教社〉では 3 用語、〈教社×社会〉では 4 用語の重複があり、コサイン類似度で見出された親子関係の類似度の高さが具体的に把握できる。また、〈教育×教社〉の重複用語は「教育」「学校」など、研究対象に関する用語が多い一方で〈教社×社会〉では、「分析」「影響」「変化」といった研究対象を捉えるための用語が多い。TF においては、重複数をみると、3 誌重複は 8 で DF に比べて少なく、重複なしの用語も〈教育〉を除いて多い。2 誌重複は〈教育×教社〉および〈教社×社会〉で多く、DF での分析と同様に、親子関係の類似度の高さを示している。具体的な用語をみると、DF で多くみられた社会科学の基礎的関心を示す用語が減少し、個別分野の特徴を示す用語が多くなっている。〈教育×教社〉では、「子ども」「教師」「生徒」といった研究対象に関する具体的な用語がとくに多い。

以上から、DF は基礎的な用語を強く反映し、社会科学の学術的文章としての共通性を取り出しやすく、TF では各雑誌の研究対象の特徴を反映しやすいと考えられる。そして、この違いが DF と TF でみるコサイン類似度の順位の違いに表れたと考えられる。つまり、教育社会学は「語り方・観点・方法」で社会学に類似し、「対象・内容」で教育学に類似していると考えられ、「教育を対象とした社会学」という教育社会学の自己規定を、実際に確認できたといえる。

3.3 頻出係り受け関係による分析

次に、各雑誌の用語をその文脈にも着目し、より詳細に分析するため、用語間の係り受けの関係の頻出上位 20 関係を表 4 に示す。用語の場合と比べると DF、TF ともに重複する数が少ない。とりわけ〈社会〉で

[3]http://code.google.com/p/mecab/

[4]取り除かれた用語は、主に数字等の一般的なストップワードおよび OCR 誤りに起因する無意味な語である。なお、OCR 誤りに起因する無意味な用語のうち、明らかに正しい用語が分かる一部の用語については人手により訂正をおこなった上で分析に利用した。

[5]http://code.google.com/p/cabocha/

表 3: 各雑誌における頻出用語 (3 誌共通:網掛、2 誌共通:外枠)

順位	DF						TF					
	教育学研究		教育社会学研究		社会学評論		教育学研究		教育社会学研究		社会学評論	
1	教育	1375	問題	793	問題	1394	教育	37952	教育	11781	問題	19289
2	問題	1367	必要	777	意味	1366	問題	19708	学校	10236	意味	15643
3	必要	1360	意味	764	必要	1359	子ども	17023	問題	9509	社会	14354
4	意味	1319	関係	737	存在	1311	学校	16348	子ども	8557	関係	12802
5	関係	1276	教育	733	関係	1282	教師	13365	研究	7582	存在	10444
6	重要	1219	研究	725	重要	1246	必要	12440	教師	7133	社会学	9630
7	研究	1215	指摘	703	指摘	1218	意味	11259	生徒	6514	必要	8754
8	指摘	1184	存在	697	社会	1182	大学	10442	意味	5577	研究	8611
9	中心	1163	社会	685	分析	1166	人間	10301	社会	5143	人間	8144
10	内容	1161	重要	685	研究	1160	研究	9926	分析	4900	日本	8083
11	目的	1152	分析	676	可能	1153	関係	8019	日本	4656	分析	7878
12	存在	1149	対象	670	対象	1131	社会	7398	必要	4638	個人	7599
13	学校	1145	学校	658	構成	1086	日本	7191	大学	4605	概念	7257
14	方法	1142	影響	649	中心	1084	自由	6971	関係	4540	家族	7082
15	課題	1129	傾向	645	形成	1068	生徒	6862	変化	4425	行為	6855
16	対象	1073	検討	642	理解	1063	課題	6792	影響	3778	理論	6724
17	展開	1064	変化	637	検討	1059	学習	6585	指摘	3488	変化	6673
18	理解	1063	中心	637	影響	1054	方法	6339	存在	3445	指摘	6394
19	発展	1055	関連	625	変化	1044	存在	6293	教育社会学	3408	人	5781
20	検討	1054	課題	606	説明	1038	内容	6103	傾向	3245	人々	5759
全体		1414		821		1436		1972409		1057558		1933685
3誌重複		11		11		11		8		8		8
2誌重複		4		7		5		7		9		4
	(×社会=1) (×教社=3)		(×教育=3) (×社会=4)		(×教社=4) (×教育=1)		(×社会=1) (×教社=6)		(×教育=6) (×社会=3)		(×教社=3) (×教育=1)	

表 4: 各雑誌における頻出係り受け関係 (3 誌共通:網掛、2 誌共通:外枠)

順位	DF						TF					
	教育学研究		教育社会学研究		社会学評論		教育学研究		教育社会学研究		社会学評論	
1	解決-問題	289	重要-意味	118	解決-問題	273	自由-教育	621	生徒-教師	351	解決-問題	423
2	教育-教育	272	生徒-教師	116	問題-問題	241	解決-問題	495	男子-女子	200	問題-問題	382
3	目的-教育	236	問題-問題	104	重要-意味	212	教育-教育	451	進学-大学	181	社会-個人	315
4	教育-問題	215	解決-問題	103	重要-問題	179	目的-教育	430	問題-問題	176	都市-農村	281
5	教育-子ども	212	調査-実施	101	重要-役割	172	発達-子ども	388	研究-教育	171	重要-意味	266
6	検討-必要	204	検討-必要	99	提起-問題	167	教育-子ども	387	職業-学歴	168	提起-問題	255
7	重要-意味	202	重要-役割	99	関連-密接	152	研究-教育	381	調査-実施	153	日本-家族	234
8	重要-課題	201	重要-課題	93	関連-問題	140	日本-教育	356	教師-学校	152	重要-役割	232
9	重要-役割	194	教育-問題	92	社会-個人	134	自由-学問	354	重要-意味	152	重要-問題	231
10	日本-教育	192	研究-教育	92	検討-必要	130	理論-実践	352	社会化-子ども	147	妻-夫	221
11	問題-問題	186	対象-分析	90	議論-展開	126	生徒-教師	350	解決-問題	147	関連-密接	210
12	研究-教育	183	重要-問題	90	関係-密接	124	機会均等-教育	342	社会-教育	146	社会学-日本	202
13	生徒-教師	176	教師-学校	83	重要-課題	120	教育-問題	335	社会学-教育	146	関連-問題	195
14	教育-内容	176	教育-学校	80	集団-個人	119	権利-保障	284	日本-教育	142	集団-個人	189
15	教育-学校	174	教育-教育	79	社会-存在	113	教師-子ども	279	教育-問題	139	議論-展開	178
16	重要-問題	171	社会-教育	79	人-人	108	問題-問題	273	家庭-学校	126	社会-人間	172
17	自由-教育	167	進学-大学	77	検討-問題	108	権利-子ども	272	教育-子ども	126	理解-意味	169
18	発達-子ども	166	規定-要因	75	関係-関係	107	検討-必要	271	対象-分析	120	社会-存在	169
19	課題-教育	161	日本-教育	74	意味-意味	107	重要-課題	263	有意-差	120	関係-関係	166
20	教師-学校	157	関係-教育	74	対象-分析	107	教育-内容	260	重要-役割	119	意味-意味	166
全体		1414		821		1436		803129		468602		776190
3誌重複		7		7		7		2		2		2
2誌重複		7		7		0		5		7		2
	(×社会=0) (×教社=7)		(×教育=7) (×社会=0)		(×教社=0) (×教育=0)		(×社会=0) (×教社=5)		(×教育=5) (×社会=2)		(×教社=2) (×教育=0)	

は、〈教育〉および〈教社〉と 2 誌重複するものが 0 になり、2 誌重複は〈教育×教社〉でのみみられる。各誌の特徴を把握しやすいのは、対象に関する用語と各誌の特徴を反映しやすい TF の係り受けである。重複のない係り受け関係を具体的に分析すると、以下のように違いが明確になる。例えば〈教育〉では、「教育」と係り受ける係り受け関係が約半数の 9 を占め、その他も「子ども」や「教師」といった教育に関する中心的な用語との係り受けが多くを占める。〈教社〉でも、やはり「教育」と係り受ける関係が多いが、「男子－女子」「職業－学歴」「調査－実施」といった係り受け関係が散見され、〈教育〉に比べると相対的に教育から関心が離れていると考えられる。また、〈社会〉では、「都市－農村」や「日本－家族」といった社会学が中心的に扱ってきた領域や、「社会－個人」や「集団－個人」といった社会学的な理論的関心を示す係り受け関係がみられる。しかし、〈教育〉や〈教社〉に比べ、研究対象を具体的に把握できる用語は多くなく、基礎的な用語が散見される。幅広い対象を扱う〈社会〉では、具体的な対象が数として分散され、対象を示す用語が上位に浮上しにくいためであると推測できる。

3.4 特定の用語に着目した分析

さらに発展的な分析として、具体的な用語 (ここでは「教育」) に着目し、その用語と係り受ける用語により分析を行う。「教育」は、〈教育〉と〈教社〉でともに TF の最頻出用語であり、この 2 誌において最も

表 5: 「教育」と係り受け関係にある用語の年代推移 (2 誌共通:網掛)

順位	全体				1950-60年代				1970-80年代				1990-00年代			
	教育学		教育社会学		教育学		教育社会学		教育学		教育社会学		教育学		教育社会学	
1	自由	621	研究	171	機会均等	163	問題	54	自由	255	地域社会	80	自由	230	日本	103
2	教育	451	社会	146	目的	147	社会学	47	子ども	206	社会学	58	教育	141	研究	94
3	目的	430	社会学	146	自由	136	工業化	43	教育	192	社会	56	子ども	139	子ども	69
4	子ども	387	日本	142	政治	135	関係	34	研究	170	研究	51	目的	133	教育	64
5	研究	381	問題	139	教育	118	社会	33	日本	166	学校	51	日本	131	社会	57
6	日本	356	子ども	126	問題	116	研究	26	目的	150	社会移動	48	公共性	123	女性	52
7	機会均等	342	教育	117	研究	107	子ども	24	問題	146	問題	46	研究	104	関係	45
8	問題	335	関係	113	機会	91	教育	21	発達	143	職業	40	権利	95	地域	43
9	内容	260	学校	107	内容	80	普及	20	学校	114	社会化	39	学校	79	歴史社会学	43
10	学校	252	地域社会	102	人間	79	関連	20	内容	114	営み	34	課題	77	学校	42
11	政治	229	社会移動	86	方法	74	機会均等	19	課題	113	関係	34	機会均等	74	社会学	41
12	発達	227	女性	85	宗教	74	必要	18	機会均等	105	子ども	33	関係	74	経済	40
13	方法	226	職業	78	質	72	経済	17	あり方	96	女性	33	問題	73	問題	39
14	課題	224	機能	70	関係	68	要求	16	方法	90	教育	32	国家	70	領域	38
15	関係	211	経済	69	結合	67	機能	15	人間	81	病理	32	目標	68	論理	37
16	権利	208	階層	69	本質	62	実践	15	権利	80	階層	31	内容	66	提供	37
17	人間	203	領域	69	日本	59	学校	14	教師	78	機能	28	発達	63	選抜	37
18	質	195	社会化	67	学校	59	機会	13	現実	78	型	28	方法	62	職業	37
19	必要	182	役割	65	生活	58	日本	13	政治	75	概念	27	あり方	62	質	34
20	あり方	178	機会均等	60	目標	58	役割	13	必要	71	機会均等	27	社会	62	家族	30
全体	803129		468602		205312		70787		312557		142519		285260		255296	
重複	8				7				6				8			

中心的な関心対象である。そこで全体および 20 年毎に、〈教育〉および〈教社〉において「教育」と係り受ける上位 20 用語を表 5 に示す。

ここには、具体的にいくつかの対比がみてとれる。ひとつは、〈教育〉では「政治」と係り受けるのにたいし、〈教社〉では「経済」と係り受けていることである。さらに、〈教育〉が「権利」「国家」など政治に関する用語と係り受け、〈教社〉は「工業化」「職業」など経済に関する用語と係り受けることをみても、教育との関係を論じる対象について、両者の棲み分けが行われてきたことが読み取れる。また、〈教育〉では「目的」や「自由」「権利」といった抽象的な用語と係り受けることが多い一方、〈教社〉では「女性」「地域社会」「職業」といった具体的な用語と係り受けることが多い。そしてこれらの用語と「教育」の関係性をみると、〈教育〉において多いのは、「教育ーのー自由」「教育ーのー権利」といったように「の」で繋がれる、教育にとってより内在的な用語である。一方〈教社〉においては、「教育ーとー経済」「教育ーとー社会移動」「教育ーとー地域社会」といったように、「と」で繋がれる教育外部／周辺的な対象との係り受け関係が多い。そしてその関係は、60 年間ほぼ一貫している。これらの分析から教育社会学が教育学との関係を念頭に、研究の対象を模索し棲み分けを図ってきたことを具体的に把握できる。

4 おわりに

本論文では教育社会学・教育学・社会学の 3 分野を対象として、論文誌に対するテキストマイニングによる分析を通して分野間の関係性を実証的に明らかにした。分析では用語・係り受け関係の TF・DF 両方を用い、雑誌間の類似度や具体的な頻度上位の用語・係り受け関係を見ることにより、各分野間の距離、関係性を様々な視点から分析した。今後の課題としては、係り受け関係を用いた分析について、現在は用語間の係り受けの方向を考慮せずに利用しているが、係り受けの方向を考慮した分析、またより詳細な言語情報を利用した分析を行うことによりまた新たな知見が得られると考えられる。さらには動詞や形容詞など用語 (名詞連続) 以外を対象とすることも考えられる。また、本論文で構築したデジタルテキストのない論文誌に対するテキストマイニング枠組を他の論文誌、特に社会科学系分野に応用し、各学問分野の特徴も検証していきたい。

参考文献

[1] 中村高康. テーマ別研究動向 (教育) : 教育社会学的平衡感覚の現在. 社会学評論, Vol. 63, No. 3, pp. 439–451, 2012.

[2] 美馬秀樹, 丹治信, 増田勝也, 太田晋. 近代文献のデジタルアーカイブ化とテキストマイニング-岩波書店「思想」を題材に. 情報処理学会研究報告. 人文科学とコンピュータ研究会報告, Vol. 2012, No. 4, pp. 1–8, 2012.

[3] 本田由紀, 齋藤崇徳, 堤孝晃, 加藤真. 日本の教育社会学の方法・教育・アイデンティティ : 制度的分析の試み. 東京大学大学院教育学研究科紀要, Vol. 52, pp. 87–116, 2012.