テキストマイニングを用いた建築情報分野の論文タイトル分析
― 日本建築学会 情報･システム･利用･技術シンポジウム論文集を対象として ―

○福田　知弘*1

キーワード：建築情報学　論文集　論文タイトル　テキストマイニング　分析

1. はじめに

2017 年，日本建築学会 情報・システム・利用・技術シンポジウムは 40 回を迎えた（但し，第 1 ～10 回である 1978～1987 年度の名称は電子計算機利用シンポジウム．1988 年度以降：現名称．以下，情報シンポ）．第 1 回が開催された 1978 年度以来，建築・都市分野の情報技術やそのシステム化，利用に関する研究論文が発表されてきた（開催は 1979 年 3 月．第 11 回の開催は年度末の 3 月に開催．第 12 回より現在までは 12 月に開催．本稿では年度表記とする）．建築分野の情報システム技術を扱う建築情報分野の発展は，論文の蓄積と表裏一体であり，情報シンポはその重要な役割を担ってきたといえよう．一方，その動向は十分に可視化されていないのではないだろうか．

学術論文は，論文タイトル，要旨，本文という枠組みで構造化されたデータといえるが，論文の内容を最も端的に表すのは論文タイトルであろう．そこで，論文タイトルに使用された単語，特に名詞（単名詞と複合名詞）を分析することにより，その学術分野の動向を把握できる可能性がある．

近年，情報技術の発達に伴い，さまざまなデータが電子データとして保存・活用されるようになってきた．紙媒体として当初刊行された媒体の電子化も進められており，情報シンポの論文集においても例外ではない．テキストマイニングは，定型化されていない文章の集まりを自然言語解析の手法を使って単語やフレーズに分割し，それらの出現頻度や相関関係を分析して有用な情報を抽出する手法やシステムを指し，学術領域を含めてさまざまな分野に適用する試みがみられる [1)-3)]．

そこで，本研究は，情報シンポで刊行されてきた論文集に掲載された論文タイトルをテキストマイニングで分析することにより，建築情報分野における動向を把握することを目的とする．

2. 方法

2.1. 分析対象

日本建築学会 情報・システム・利用・技術シンポジウム論文集（但し，1979～1988 年：電子計算機利用シンポジウム論文集，1989 年以降：現名称．以下，情報シンポ論文集）に掲載された論文タイトル（主題と副題を含む）を対象とした．論文数は 2628 編である．

分析用ソフトウェアとしては，自然言語処理と多変量解析の両方を実行可能なフリーソフトウェアである KH Coder を用いた [4)5)]．

2.2. データの前処理

(1) 論文タイトルデータの整備

まず，情報シンポ論文集は年代ごとに以下の保存状態であり，テキストデータ化するための作業を実施した．

- 1978-2009 年度：ラスター形式で電子化されている．そのため，筆者が OCR（optical character reader）により処理を行い，主題・副題をテキストデータ化した．OCR による変換エラーは手入力で修正した．
- 2010-2011 年度：紙媒体で保存されている．そのため，筆者がラスター化を行った上で OCR により主題・副題をデータ化した．
- 2012-2017 年度：ベクター形式で電子化されている．主題・副題のテキストデータをそのまま利用した．

建築情報分野で扱う内容は年代ごとに異なることが予想される．そのため，1978 年度から 2017 年度までの 40 年間を 10 年度ごとに 4 期に分割し，それぞれ「1 期（1978-1987 年度）」「2 期（1988-1997 年度）」「3 期（1998-2007 年度）」「4 期（2008-2017 年度）」とした．

次に，KH Coder には，分析対象のテキストとその外部変数とのリスト（Excel 形式）をテキスト形式に自動変換して，外部変数と共に入力する．そのため，本研究で分析するために，情報シンポ論文集に収録された全ての論文タイトルを，論文タイトルに 4 期の区別および出版年が紐づけられたリスト形式として整備した（以下，論文タイトルリスト）．尚，同一年で出版された連続論文の主題（または副題）は重複するため，最初の一編以外は削除した．

(2) 強制抽出する語と使用しない語の指定

KH Coder では，日本語データから単語を取り出すために，茶筌という外部プログラムを使用して形態素解析を行う．しかし，デフォルト設定による形態素解析の抽出では，一般的な単語に限られているために，建築情報分野に関する用語が抽出されないか，抽出されたとしても下位に位置づけられてしまい研究動向の把握に困難が予想される．また，品詞単位に細かく分割されてしまう場合があることが予想される．例えば「建築設計」という用語は「建築」と「設計」という 2 つの語として認識され，複合名詞の抽出が行えない．そこで，強制抽出する語を指定するために，

以下の作業を実施した.

まず，KH Coder では，専門的な複合語を抽出するために， TermExtract を利用して「複合語の検出」機能を利用できる．本研究では，この機能を用いて，論文タイトルリストより，複合語を抽出した．結果，5041 の複合語が抽出され，最高スコアは 6489.4（建築構造物），最低は 1（WWW ブラウザなど 12 語），平均は 54.6 であった．スコアが 100 以上の複合語 474 語のうち，議論の実質に関係しない 1 語（表 1）を除く 473 語を「強制抽出する語」として指定した.

次に，建築分野の専門用語が収録された「建築学用語辞典」[6]の見出し語の内, 2 語以上の複合語 19798 語のうち，予備調査段階で単語の一部などに誤検出された 12 語（表 1）を除く 19786 語を「強制抽出する語」として指定した．また，情報技術分野の専門用語が収録された「IT 用語辞典 e-Words」[7]の見出し語 9999 語のうち，予備調査段階で誤検出された 67 語（表 1）を除く 9932 語を「強制抽出する語」として指定した．さらに，同義語が存在する 32 語を一つの複合語にそれぞれ整理し,「強制抽出する語」として指定した（表 2）.

最後に，茶筌で抽出され上述の 4 作業で抽出しなかった建築情報関連用語 32 語を指定した[注].

(3) 前処理の実行

上述した指定作業の後，KH Coder で「前処理の実行」を行った．分析対象ファイルの文章から語を切り出し，その結果をデータベース化する．分析に使用する語は，上記(2)項で指定した名詞 30255 語（KH Coder 上では「タグ」と表示）のみとした．抽出された語数を表 3 に示す.

2.3. 分析期間

全期間，および，2.2(1)で述べた 4 期別とした.

3. 結果

3.1. 全期間

全期間での抽出語の合計は, 10978 語(表 3(B)の 47%)，異なり語数で 2014 語（表 3（D）の 46%）であった．上位 100 位までの頻度順位を表 4 に示す.「システム」「開発」が圧倒的に多く，100 頻度以上の名詞は「モデル」「3D」「評価」「型」「建築」「情報」「設計」「シミュレーション」「データ」と続いた.

次に，共起ネットワーク図を作成した（図 1）．共起ネットワーク図は，抽出語をノードとして，共起関係（出現パターンの似通ったもの）を線で結んだ図である．本研究では，名詞の出現頻度に応じてノード円のサイズが変化し，Jacaard 係数で測定した共起の度合いに応じて共起関係を表す線の太さが変化するように表現した．年度の経過による共起関係の変化を観察するために，期間を見出しとして描いた.「描画する共起関係」の閾値は 120 と設定した.

結果，各期間でのみ特徴的な名詞を Degree 1，2 つの期間に共通の名詞を Degree 2, 3 つの期間に共通の名詞を Degree 3，全期間に共通の名詞を Degree 4 に分類できた．上述した出現頻度 11 位以内の名詞は「設計」を除いて Degree 4 に分類されている．これらは，全期間に共通した建築情報分野に不変の名詞といえる.

また，Degree 3 のうち「建物」「都市」「情報」は 4 期に含まれた．Degree 2 のうち 3 期と 4 期に含まれた名詞は「形態」「地域」「空間」「環境」「制御」「最適化」「デザイン」「画像」「遺伝的アルゴリズム」であった．これらは, Degree 1 で 4 期に含まれた名詞「アルゴリズム」「BIM」「ロボット」「地震」「災害」「オフィス」「形状」「モニタリング」「知的活動」「拡張現実感」と併せて，現在の中心的な名詞といえる.

3.2. 期間別

期間別の共起ネットワーク図を図 2～4 に（「描画する共

表 1. 「強制抽出する語」に含めない除外語

TermExtract	基礎的研究
建築学用語辞典	ET, IR, LL, PA, PE, PS, SS, VE, クロ, ネット, ラム, ループ
IT 用語辞典 e-Words	AC, AD, AI, AM, AND, AP, AR, AS, au, BI, BL, CA, CE, CI, COM, CT, DA, DES, DI, EA, EC, ED, FA, FC, FE, FM, GA, IA, IC, ID, IM, IT, LD, LL, LT, MA, MO, NC, NI, NS, OR, OS, PL, PO, PS, RA, RO, RT, SA, SD, SE, sh, SI, SP, SS, ST, TA, TS, U, UD, VE, vi, タグ, ナル, パス, ランド, ループ

表 2. 同義語の変換

使用語	同義語
3D	3 次元, 三次元
BIM	Building Information Modeling
CG	コンピュータグラフィックス, C.G., コンピュータ・グラフィックス
EWS	エンジニアリングワークステーション
FM	ファシリティマネジメント, ファシリティ・マネジメント
VR	バーチャルリアリティ
Web	WWW, ウェブ
アニメーション	アニメイション
遺伝的アルゴリズム	GA
拡張現実感	AR
グラフィック・ディスプレイ	グラフィックディスプレイ, Graphic Display
ケース・スタディ	ケーススタディ, ケーススタデイ, ケース・スタデイ
コンピュータ	コンピューター
シミュレーション・システム	シミュレーションシステム
シミュレーション・プログラム	シミュレーションプログラム
シミュレータ	シミュレーター
情報技術	IT, インフォメーションテクノロジー
人工知能	AI, A.I.
データベース	DB
テクノロジ	テクノロジー
デザイン・ツール	デザインツール, 設計ツール
鉄筋コンクリート	RC
パソコン	パーソナル・コンピューター, パーソナル・コンピュータ, パーソナルコンピューター, パーソナルコンピュータ, PC ※
プレストレスト・コンクリート	プレストレストコンクリート, PC ※
ボリューム	ヴォリューム
マイクロ・コンピュータ	マイクロコンピュータ
マルチ・エージェント・システム	マルチエージェントシステム
メッシュ・データ	メッシュデータ
モデラ	モデラー
レイ・トレーシング	レイトレーシング
レイヤ	レイヤー
無線 IC タグ	RFID

表 3. 抽出された語数

対象	論文数	総抽出語数(A)	使用語数(B)	異なり語数(C)	使用語数(D)
全て	2628	37255	23208	4703	4366
1 期	688	7795	4815	1733	1561
2 期	750	9507	5946	2062	1873
3 期	543	8833	5494	1831	1675
4 期	647	11120	6953	2310	2124

起関係」の閾値は 60），頻度別抽出語の上位 30 位までを表 5 に示す．表 5 には，4 期の上位 30 位以内の抽出語に

表 4. 全期間における頻度別抽出語（上位 100）

順位	抽出語	頻度	順位	抽出語	頻度
1	システム	382	49	最適	31
2	開発	310	49	自動	31
3	モデル	165	53	プロセス	30
4	3D	140	53	業務	30
5	評価	135	53	性能	30
6	型	131	53	知識	30
7	建築	124	53	部材	30
8	情報	112	58	ツール	29
9	設計	107	58	ベース	29
10	シミュレーション	102	60	立体	27
11	データ	100	61	CAD システム	26
12	空間	88	61	インターネット	26
13	都市	81	61	コミュニケーション	26
14	デザイン	76	61	大学	26
15	建物	73	61	値	26
16	プログラム	69	61	避難	26
17	構造物	68	67	マイクロ・コンピュータ	25
18	遺伝的アルゴリズム	67	67	可能性	25
19	CG	66	67	市街地	25
20	コンピュータ	64	70	モデル化	24
21	環境	63	70	モニタリング	24
22	最適化	61	70	可視化	24
23	パソコン	58	70	環境デザイン	24
24	データベース	56	70	社会	24
25	制御	55	70	鉄骨	24
26	CAD	53	76	エージェント	23
27	ネットワーク	51	76	モデリング	23
27	形態	51	78	都市空間	22
29	アルゴリズム	50	78	平面	22
30	構造	46	80	VR	21
31	地域	45	80	センサ	21
32	地盤	43	80	デジタル	21
33	Web	42	80	遠隔	21
33	景観	42	80	建築設計	21
33	支援システム	42	80	入力	21
36	建築物	41	80	列	21
37	画像	40	87	ビル	20
38	地震	39	87	応答	20
39	建築構造物	38	89	工法	19
40	GIS	37	89	構造設計システム	19
41	BIM	36	89	骨組	19
42	オフィス	35	89	地震応答	19
42	ロボット	35	89	都市景観	19
42	管理	35	94	建築計画	18
42	住宅	35	94	構造解析	18
46	鉄筋コンクリート	34	94	室内	18
47	形状	33	94	状況	18
48	知的	32	94	人間	18
49	イメージ	31	94	図面	18
49	ニューラルネットワーク	31	94	地図	18

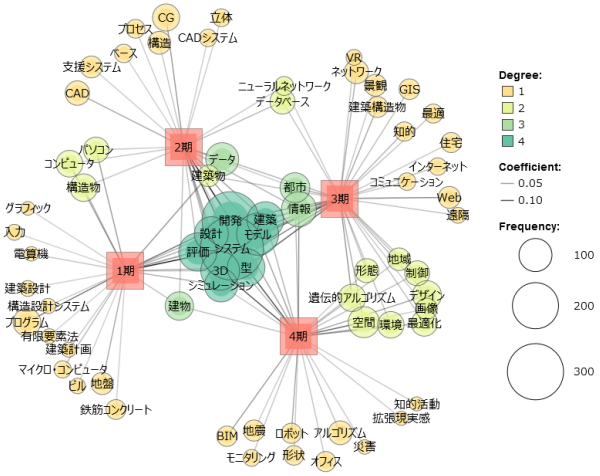

図 1 期間を見出しとした共起ネットワーク図

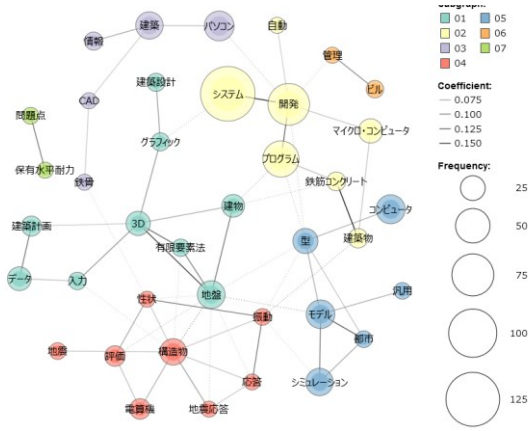

図 2 第 1 期共起ネットワーク図

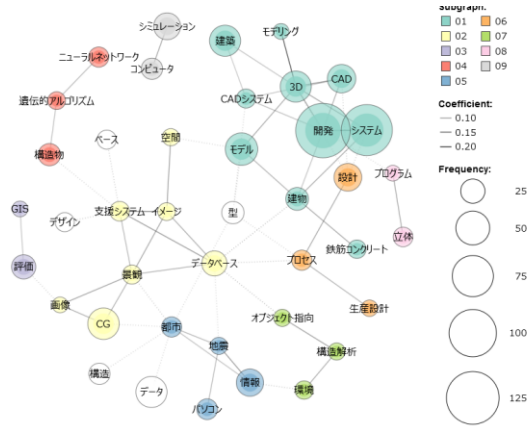

図 3 第 2 期共起ネットワーク図

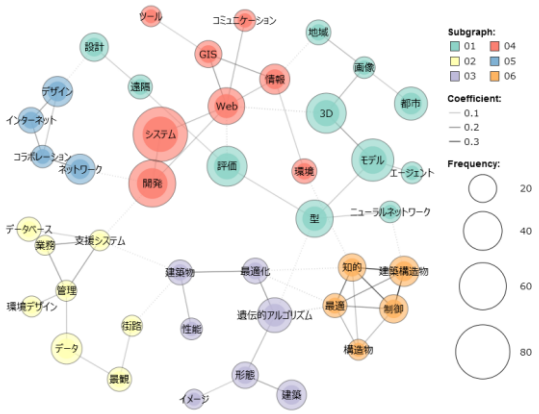

図 4 第 3 期共起ネットワーク図

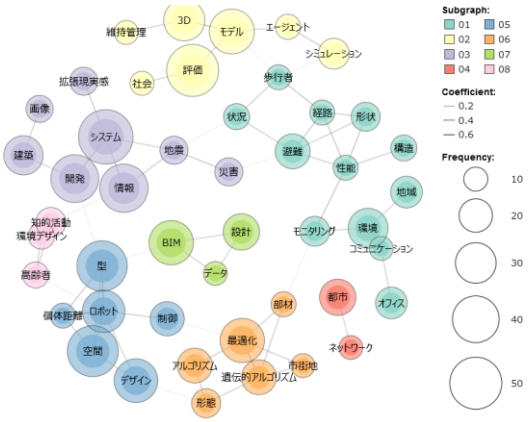

図 5 第 4 期共起ネットワーク図

表 5. 頻度別抽出語（期間別．太字は同数）

順位	1 期 1978-1987 年度 抽出語	頻	2 期 1988-1997 年度 抽出語	頻	3 期 1998-2007 年度 抽出語	頻	4 期 2008-2017 年度 抽出語	頻	順位変化 1→2→3→4 期
1	システム	129	開発	134	システム	81	システム	55	1→2→1→1
2	開発	74	システム	117	開発	60	評価	50	18→13→4→2
3	プログラム	55	モデル	46	モデル	48	空間	48	118→36→11→3
4	パソコン	37	CG	43	3D	**43**	型	47	12→15→6→4
5	コンピュータ	35	3D	**42**	評価	**43**	情報	43	23→10→11→5
6	モデル	34	データ	**42**	型	37	開発	42	2→1→2→6
7	地盤	32	建築	**42**	Web	36	モデル	37	6→3→3→7
8	シミュレーション	**31**	CAD	35	遺伝的アルゴリズム	32	BIM	**36**	NA→NA→NA→8
9	建築	**31**	設計	33	都市	29	最適化	**36**	373→74→23→8
10	構造物	**29**	情報	32	データ	25	デザイン	35	228→25→11→10
11	設計	**29**	シミュレーション	31	デザイン	**24**	ロボット	33	228→NA→NA→11
12	型	26	データベース	27	ネットワーク	**24**	3D	31	13→5→4→12
13	3D	**24**	評価	25	空間	**24**	環境	29	118→29→25→13
14	マイクロ・コンピュータ	**24**	建物	22	建築構造物	**24**	建築	28	8→5→16→14
15	データ	23	型	**21**	情報	**24**	設計	**24**	10→9→18→15
16	建物	21	構造物	**21**	シミュレーション	**23**	都市	**24**	30→20→9→15
17	電算機	**17**	コンピュータ	**19**	建築	**23**	アルゴリズム	**22**	90→56→28→17
18	評価	**17**	構造	**19**	制御	**21**	遺伝的アルゴリズム	**22**	NA→33→8→17
19	建築計画	16	パソコン	18	設計	**21**	建物	**22**	16→14→59→17
20	建築物	**15**	ニューラルネットワーク	**17**	知的	20	避難	**22**	158→457→NA→17
21	構造設計システム	**15**	プロセス	**17**	GIS	19	制御	21	158→42→18→21
22	鉄筋コンクリート	14	都市	**17**	形態	18	オフィス	**18**	373→56→59→22
23	グラフィック	**13**	支援システム	**16**	インターネット	**17**	地域	**18**	90→84→25→22
24	建築設計	**13**	立体	**16**	最適化	**17**	シミュレーション	17	8→11→16→24
25	情報	**13**	CAD システム	**15**	環境	**16**	モニタリング	**16**	NA→281→83→25
26	ビル	**12**	デザイン	**15**	建築物	**16**	知的活動	**16**	NA→NA→NA→25
27	入力	**12**	ベース	**15**	地域	**16**	形状	**15**	52→108→104→27
28	有限要素法	**12**	景観	**15**	アルゴリズム	**15**	形態	**15**	44→48→22→27
29	CAD システム	**11**	環境	**14**	データベース	**15**	拡張現実感	**14**	NA→457→419→29
30	振動	**11**	自動	**14**	遠隔	**15**	災害	**14**	NA→457→245→29

対して，期間ごとの順位変化も示す．

1 期（1978-1987 年度）：構造計算系の名詞が特徴的である．現在（4 期）では「マイクロ・コンピュータ」「電算機」「パソコン」などは出現しなくなった名詞，「プログラム」「コンピュータ」「グラフィック」などは低頻度の名詞となった．

2 期（1988-1997 年度）：「CG」「CAD」「データベース」「ニューラルネットワーク」など技術を表す名詞，「都市」「景観」など建築情報分野が扱う対象の拡がりを示す名詞，「プロセス」「支援システム」などの名詞が新しく特徴的である．一方，現在では「CAD」「CG」「データベース」「ニューラルネットワーク」「プロセス」などは低頻度である．

3 期（1998-2007 年度）：「Web」「インターネット」「ネットワーク」「コラボレーション」「遠隔」などのインターネット関連技術・手法や「遺伝的アルゴリズム」「知的」などの人工知能関連の名詞，「環境」「地域」など対象の拡がりを示す名詞，「制御」「最適化」「GIS」「VR」などの名詞が特徴的である．現在では「コラボレーション」は出現しなくなった名詞，「Web」「インターネット」「遠隔」などは低頻度の名詞となった一方，「知的」は「知的活動」「知的照明システム」への拡張が見られる．

4 期（2008-2017 年度）：「BIM」「ロボット」「拡張現実感」などの新たな技術，「オフィス」「避難」「災害」「モニタリング」「知的活動」など対象の拡がりを示す名詞が特徴的である．また，「評価」「空間」「最適化」「環境」「アルゴリズム」は増加傾向である．

4. まとめ

本研究では，建築情報分野における研究動向を把握するために，過去 40 年間に発刊された情報シンポ論文集の全論文タイトルに着目して，全期間と 4 期間それぞれにおいて，テキストマイニングによる分析を実施した．建築情報分野の名詞の抽出のために，建築分野，情報技術分野の専門用語等からなる建築情報分野の名詞リストを作成した上で分析した．結果，建築情報分野の学術用語として不変の名詞を抽出と共に，4 期間毎に特徴的な名詞や名詞の出現頻度の順位変動を抽出することができた．

建築情報分野の用語リストの作成と更新のあり方の議論の必要性を今後の課題としたい．

注

IFC，アクチュエータ，遠隔，オントロジー，可視，画像，規模，業務，形状，形態，経路，コミュニケーション，コラボレーション，最適，室内，自動，社会，写真，状況，性状，センシング，知識，知的，人間，ビル，ファジイ，フラクタル，平面，モーション，リアルタイム，立体，ワークショップの 32 語

謝辞

1978 年度以降，情報・システム・利用・技術シンポジウム（旧電子計算機利用シンポジウム）論文集の刊行，並びに，その電子化に携わった関係各位に深甚の謝意を表する．

[参考文献]

1) 佐久嶋研, 佐々木秀直, 田代邦雄, 2012, テキストマイニングを用いた学会誌論文タイトルの時系列分析―日本神経学会誌「臨床神経学」の分析―, 医療情報学, 32(6), 315-321.
2) 木原大志, 小立雄大, 緒方雄基, 柳信栄, 小林祐司, 2016, テキストマイニングを用いた災害情報の抽出と分析―地方紙 HP が熊本地震をどう伝えたか―, 日本建築学会 第 39 回情報・システム・利用・技術シンポジウム 2016 (報告), 149-152.
3) 松岡広, 森川想, 2017, テキスト分析による審議会等の議事の可視化に向けて―在宅医療に関する中央社会保険医療協議会の議事を例に―, 社会技術研究論文集, 14, 73-83.
4) 樋口耕一, 2004, テキスト型データの計量的分析―2 つのアプローチの峻別と統合―, 理論と方法, 19(1), 101-115.
5) KH Coder: 計量テキスト分析・テキストマイニングのためのフリーソフトウェア, http://khcoder.net/ (参照 2018-10-3)
6) 日本建築学会, 2008, 建築学用語辞典 第 2 版, 岩波書店
7) IT 用語辞典 e-Words, http://e-words.jp/ (参照 2018-10-3)

＊1 大阪大学 大学院工学研究科 環境・エネルギー工学専攻 准教授 博士(工学)