研究論文

テキストマイニングによる傾向・様相の分析

稲田　愛（岡山理科大学大学院マネジメント研究科）
森　裕一（岡山理科大学経営学部）

要約：

本研究では，作者の文章作成の傾向や作品の内容など，対象の傾向や様相を文章データを基に把握することを目的に，ニュース時事能力検定の模擬問題サイトの全問題文と劇場版アニメシリーズの映画4作品のレビュー文を事例として計量テキスト分析を行った。分析では，テキストマイニング手法として提供される特徴分析，共起分析，対応分析を用いた。その結果，検定問題における出題傾向や出題者の作問の特徴，映画各作品の概要や評価スコアとの関連を明らかにすることができ，判断や行動を起こすための計量テキスト分析による情報縮約の機能が十分に働くことが確認できた。

キーワード：

KH Coder，スクレイピング，出題傾向，レビュー分析，特徴分析，共起ネットワーク，対応分析

１．はじめに

インターネットの利用拡大や技術の発展によって，さまざまな情報が大量に飛び交うようになっている。このような状況下で，限られた時間内に，効率的に情報の取捨選択を行い，判断や行動を起こすには，当該の情報の傾向や様相を素早く，また，要点を絞って情報を集約することが必要である。しかし，たとえば，情報が文章で提供されるものは，「読む」ことが必要であるため，瞬時に傾向や内容を把握することがむずかしい。特に，ある作者が記した多くの文章からその作者の傾向を知りたいときや，ある対象を表現したたくさんの文章からその対象の様相を知りたいときなどである。具体的には，テストの長文問題の出題傾向，歴史上の書物や文書の作者の特定，口コミサイトやレビューサイトから利用するホテルや飲食店を決めるような場面である。

これに対して，文章データ（テキストデータ）を構成する単語の頻度や係り受け関係を統計的に分析して，対象全体の傾向や文章に割り当てられた属性との関係を明らかにできるテキストマイニングの手法を適用することが考えられる。これらを利用すれば，作者ごとの文章作成の傾向を把握したり，口コミサイトやレビューサイトから対象商品や作品の特徴などを把握したりすることが可能となる。

そこで，本研究では，先にあげた複数の場面から2つの問題場面を想定して，膨大なテキストの情報をテキストマイニングにより，傾向や様相がどのように把握できるか，そのことから計量テキスト分析の情報縮約の機能が十分に働いているかを確認する。1つは検定問題の出題傾向を見ること，もう1つは映画のレビュー文から作品の内容を把握することである。前者に対しては，ニュース時事能力検定の模擬問題を提供する「めざせニュース検定」（https://www.soci.ous.ac.jp/newskentei/index.html）の問題文を用い，出題および出題者の傾向を明らかにする。このデータは，各問題の出題者がはっきりしているため，出題者による出題傾向を確認できる典型的なデータである。後者では，劇場版「名探偵コナン」の映画のレビューサイト（https://filmarks.com/）のレビュー文を対象として，名探偵コナンの各作品の特徴や評価スコアとの関連を明らかにする。この映画のレビューデータは，見どころやストーリーなどが自由記述で書かれた大量のテキストからなり，複雑な文章から特徴を取り出せるかを確認することに適したデータである。

ここで，問題となってくるのは，分析すべき文章をどのように効率的に集めるかである。たとえば，レビューサイトでは1つの対象に対するレビューが1000以上もあることがある。テキストの収集に時間がかかれば，テキストマイニングによる情報把握に実効性が期待できない。これに対して，スクレイピング技術を使ってWebから自動収集することが考えられる（たとえば，小川（2022）など）。「めざせニュース検定」の問題文すべては手元にあるので，映画「名探偵コナン」のレビューサイトからのレビュー文の収集にスクレイピングによる自動収集を実施する。

以下，2 節では，本研究で用いるテキストマイニングの手法について概説し，3 節ではニュース検定の出題の分析，4 節では名探偵コナンの映画のレビュー文の分析を行い，最後にまとめを行う。

２．テキストマイニングについて

テキストマイニングとは，文章（テキスト）データと採掘するという意味のマイニング

を組み合わせた用語で，大量の文章データから有益な情報を取り出す手法の総称である。

テキストマイニングの主な分析手法として，検索機能，頻度分析，特徴分析，共起分析，時系列分析，文書分類・文書クラスタリング，評判分析などがある（上田 他，2008；金，2009）。ここでは，本研究で使用する特徴分析，共起分析，対応分析について概要を示す。

- 特徴分析：文章データに対して偏って多く出現している単語をその文章に対する特徴的な単語として抽出する手法のことである。特徴的な単語を特定するため，単語の出現の偏り具合を数値化する。その数値を「特徴度」とよぶ。
- 共起分析：同じ文章の中で，単語が共に出現することを「共起」とよぶ。単純な頻度ではなく，単語と単語の共起，単語と属性の共起，データと単語の共起など，さまざまなパターンを把握しようとするものである。この分析手法のことを共起分析とよぶ。テキストにおいて，単語の共起パターンは重要な情報であり，この関係を把握するための統計的な手法が共起ネットワークで，そのネットワークを可視化したものが共起ネットワーク図である。この図では，共起関係の結びつきの強さも表現されるため，文章群の内容の把握が容易となる。
- 対応分析：単語間の関係性を散布図で視覚的に表現する方法のことである。多くの文章に出現する単語は原点近くに配置され，特定の文章に偏って出現する単語は原点から離れた場所に配置される。また，互いに関連の強い単語は，原点から同じ方向に配置される。外部変数（各文章の属性）とのクロス集計を対応分析すると，その属性と単語の関係から，各外部変数の特徴づけを行うことができる。

テキストマイニングを行うための解析ツールとして，フリーソフトウェアのKH Coder（樋口 他，2022）が有名である。本研究でも，これを用いて実際の分析を行う。

テキストマイニングを利用した先行研究として，出題傾向を計量テキスト分析した事例として中島・早田（2019）がある。また，レビュー文の分析では小川（2022）がある。中島・早田（2019）では，問題の分野ごとに抽出語と共起関係を明らかにし，各分野において求められる思考力の傾向を確認している。小川（2022）では，スクレイピングによるレビュー文と関連情報を収集し，映画作品の評価に大きな影響を及ぼす要素の抽出と構造の把握，2つの映画の要因分析より映画評価が高まる共通的な要因の抽出を行っている。なお，いずれも頻度分析と共起分析が中心とした分析であるが，本研究では，これらに加え，対応分析も行い，外部変数による特徴づけも行う。

３．ニュース検定問題の分析

３．１．データ

多くの文章から，そこでよく取り上げられているテーマや作者の傾向を把握することを考える。具体的には，検定試験の出題傾向を知るという場面を想定し，ニュース時事能力検定の模擬問題を提供する「めざせニュース検定」の出題文を分析する。

「ニュース時事能力検定（ニュース検定，N検）」とは，新聞やテレビのニュース報道を読み解き，活用する力（時事力）を養い，認定する検定である（日本ニュース時事能力検定協会，https://www.newskentei.jp/ a_index.html）。この対策サイトとして，岡山理科大学総合情報学部社会情報学科（のち，経営学部）が2008年10月2日から始めた「めざせニュース検定」（社会情報学科・経営学部，https://www.soci.ous.ac.jp/newskentei/index.html，毎日新聞岡山版に週１回掲載）で提供される模擬問題全621問のうち，学生が作成した9問を省く612問を使用する。それぞれの問題から「問題文」，「正解番号」を取り出し，これに「問題の出題者」（森，八木，木村）を加えたデータをテキストマイニングする。

KH Coder上での前処理として，品詞による語の取捨選択を行う。分析対象を「名詞」，「サ変名詞」，「形容動詞」，「固有名詞」，「組織名」，「人名」，「地名」，「ナイ形容」，「タグ」，「動詞」，「形容詞」，「名詞C」に限定した。さらに，強制抽出する語として，「厚生労働省」，「経済協力開発機構」を指定し，使用しない語として，「次」，「月」，「年」を指定した。

３．２．共起ネットワークによる全問題の傾向

問題文612問に対して共起分析を行って作成されたネットワーク図が図3.1である。これより，「厚生労働省」，「調査」，「人口」，「割合」のネットワークから，文章中では「厚生労働省が発表した調査では」，「人口に占める割合は」といった使われ方をしており，「統計」に関する話題が表れていること，「オリンピック」，「開催」，「東京」，「国際」のネットワークから，文章中では「東京オリンピックの開催に合わせて」，「国際オリンピック」といった使われ方をしており，「オリンピック」に関する話題が表れていること，「遺産」，「世界」，「登録」，「決まる」のネットワークから，文章中では「世界文化遺産登録を決めた」といった使われ方をしており，「遺産」に関する話題が表れていることが読み取れる。

３．３．対応分析による出題者の作問の特徴

抽出語（問題文を構成する単語）と外部変数（出題者）を組み合わせた対応分析を行って得られる散布図が図3.2である。

原点の左上方向に，出題者「森」があり，その方向に，「協定」，「改正」，「成立」，「女子」，「日本人」などの語が出てきており，「森」が出題した問題の特徴的な語となっている。左下側には，出題者「八木」と「首相」，「都道府県」，「名称」，「女性」，「平均」などの語が出てきており，「八木」が出題した問題の特徴的な語だといえる。右側には，出題者「木村」と「条約」，「記述」，「登録」，「社会」，「地域」などの語が原点から離れており，出題者「木村」が出題した問題の特徴的な語だといえる。出題者によって頻出する語が異なるということ，すなわち，全員が似たような分野から出題しているわけではなく，それぞれ明瞭に分かれていることがわかった。このことから，テストの問題文や歴史上の人物が書いた手記，レビューの自由記述やホテルやお店の口コミの分析に利用できることが示唆される。

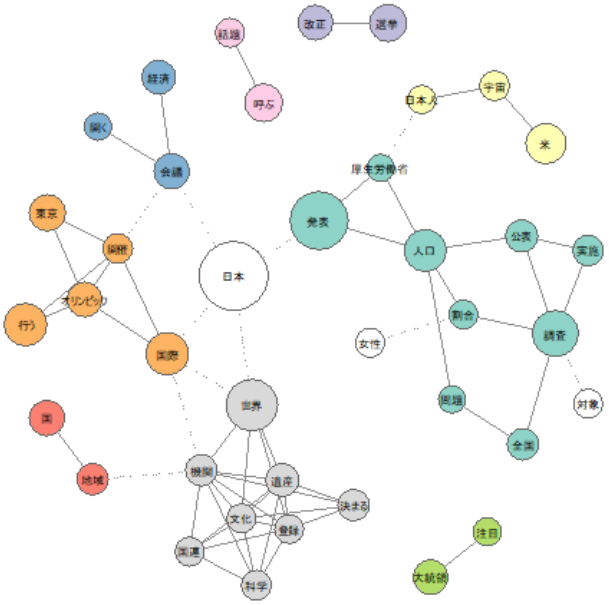

図3.1　共起ネットワーク図

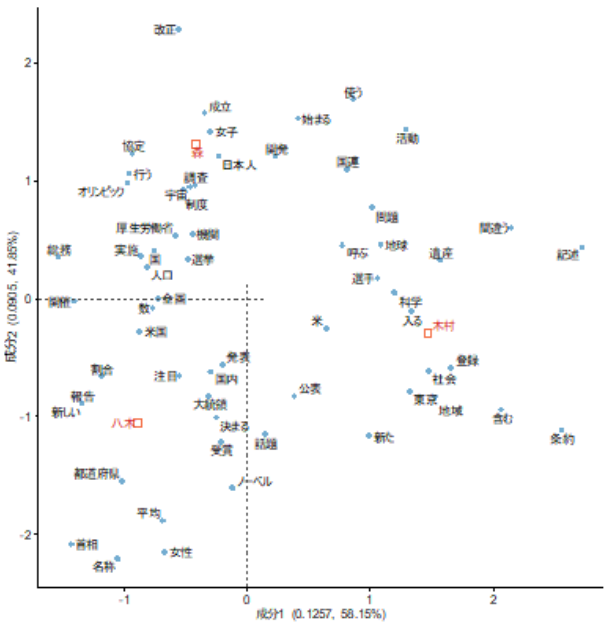

図3.2　対応分析

４．映画「名探偵コナン」のレビューの分析

４．１．データ

ある対象を表現したたくさんの文章からその対象の様相を把握することを考える。具体的には，映画のレビューサイトのレビュー文を分析する。映画は，劇場版「名探偵コナン」を用いる。

レビューデータは，映画のレビューサイトである「Filmarks 映画」（株式会社つみき，https://filmarks.com/）より，スクレイピングによって収集する。スクレイピングとは，ネット上のサイトの構造を解析して，必要なデータを（指定した条件に合う文字列をすべて）自動的に取り出し，集めてくることである（森，2019；清水，https://ai-inter1.com/python-webscraping/；清水，2022）。このスクレイピングによって，劇場版「名探偵コナン」の1作目から25作目までの全25作のレビュー文と評価の星（★）の数（スコア）を収集する。収集は2023年1月19日から1月30日の期間に行った。

収集した25作品のうち，次の4つの作品を分析する。本研究はストーリーの把握を目的としているので，大きくストーリーが変わる作品として選んだものである。劇場版の1作目である「時計じかけの摩天楼」，主人公である江戸川コナンの天敵である怪盗キッドが初登場する「世紀末の魔術師」，謎の集団である黒の組織が初登場する作品である「天国へのカウントダウン」，本論文執筆時点での最新作である「ハロウィンの花嫁」である。各作品のスコアは，小数点第一位で四捨五入し整数化しておく。なお，評価をつけていないレビューについては，スコアなしとしてカウントしている。4作品のスコアごとのレビュー数とレビューが記載された期間を表4.1に示す。

「ニュース検定」と同様，前処理として，「名詞」，「サ変名詞」，「形容動詞」，「固有名詞」，「人名」，「形容詞」を分析対象とする。また，作品ごとに強制抽出する語を加えておく。

表4.1　各作品のスコアごとのレビュー数

タイトル	★の数							記載期間
	5	4	3	2	1	※	計	
時計じかけの摩天楼	237	1588	559	28	2	253	2667	13/4/4～23/1/15
世紀末の魔術師	224	1372	429	24	4	200	2253	13/5/19～23/1/13
天国へのカウントダウン	275	1495	379	18	2	245	2414	13/2/15～23/1/13
ハロウィンの花嫁	1509	4797	688	62	18	811	7885	21/12/4～23/1/24

※　スコアなし

４．２．特徴語による分析

各作品の出現数上位30件の単語を抽出したものを表4.2～4.5に示す。これらより，次のことが観察される。

(1) 時計じかけの摩天楼

表4.2に示された偏って多く出現している特徴的な単語より，「爆弾」，「爆発」，「爆破」からストーリーに爆発があること，「好き」，「面白い」，「良い」から高評価を得ている映画であることがわかる。また，「赤い」から赤色の何かがキーとなっていること，「電車」から事件に電車が関わってくること，「コナン」，「新一」，「蘭」，「森谷」があり，この作品のキーパーソンであることがわかる。実際，「森谷」はこの映画から登場する人物であり特徴的な人物であることが抽出されている。

(2) 世紀末の魔術師

表4.3に示された特徴的な単語より，「キッド」という存在が最も多く出てきており，重要な存在であること，「ロマノフ王朝」，「歴史」からロマノフ王朝や歴史に関するストーリーであることがわかる。「怪盗」という存在が出てくること，「正体」から何かの正体について触れられるストーリーがあることがわかる。「ラスト」，「仕掛け」からラストに何かあること，何かしらの仕掛けが施されていることがわかる。

(3) 天国へのカウントダウン

表4.4より，「アクション」，「脱出」からアクションシーンや脱出シーンが多いこと，「組織」から何かの組織がキーになること，「ビル」，「ツインタワー」からこれらの建物が関係していることがわかる。「灰原」，「元太」，「歩美」，「光彦」からこれらの人物がキーパーソンであること，「可愛い」から可愛いものが登場することがわかる。

(4) ハロウィンの花嫁

表4.5より，「学校」から学校がストーリーに関わってくること，「アクション」からアクションシーンが多くあること，「警察」，「刑事」から警察官が重要な存在であることがわかる。また，「安室」，「松田」，「降谷」，「佐藤」，「高木」という人物がキーパーソンであることがわかる。

４．３．共起ネットワークによる作品の様相

各作品の共起ネットワーク図を図4.1～4.4に示す。共起ネットワーク分析を行うにあたって，登場人物名は作品の説明には冗長となるので，「人名」を抜いたものを使用する。

表4.2「時計じかけの摩天楼」の抽出語

抽出語	出現数	抽出語	出現数
コナン	2609	新一	408
映画	1373	蘭	382
爆弾	856	推理	312
犯人	762	感じ	276
作品	680	アニメ	275
劇場	664	動機	265
好き	607	ラスト	264
最後	561	爆発	260
面白い	557	爆破	245
シーン	549	森谷	240
赤い	516	シリーズ	232
笑	462	電車	219
探偵	446	鑑賞	212
良い	414	展開	209
事件	412	伏線	205

表4.3「世紀末の魔術師」の抽出語

抽出語	出現数	抽出語	出現数
キッド	2192	平次	269
コナン	1975	ストーリー	268
映画	1006	話	217
怪盗	726	ラスト	207
好き	589	仕掛け	188
作品	579	感じ	176
登場	487	シリーズ	154
面白い	480	ロマノフ王朝	154
笑	414	最高	150
劇場	407	素敵	142
シーン	386	歴史	138
最後	386	多い	135
良い	371	事件	132
探偵	275	正体	131
犯人	271	推理	128

表4.4「天国へのカウントダウン」の抽出語

抽出語	出現数	抽出語	出現数
コナン	2112	元太	358
映画	991	歩美	349
探偵	951	最後	346
少年	717	活躍	343
シーン	662	あゆみ	341
ビル	651	犯人	339
組織	649	光彦	330
灰原	647	事件	305
好き	622	可愛い	294
作品	574	脱出	279
笑	480	ツインタワー	248
アクション	453	蘭	247
良い	416	動機	211
劇場	400	ラスト	200
面白い	396	感じ	196

表4.5「ハロウィンの花嫁」の抽出語

抽出語	出現数	抽出語	出現数
コナン	8231	作品	1249
映画	4590	佐藤	1239
警察	3190	爆弾	1228
面白い	3124	話	1188
安室	3023	最後	1186
良い	2887	感じ	1151
学校	2828	ストーリー	1146
刑事	2648	劇場	1115
シーン	1957	最高	1089
犯人	1751	探偵	959
笑	1741	事件	923
アクション	1503	降谷	872
松田	1496	多い	854
好き	1423	映画館	830
高木	1265	展開	801

(1) 時計じかけの摩天楼

図4.1より，「事件」，「爆発」，「犯人」を中心とするネットワークから犯人が起こす事件は爆発が起こるストーリーであること，「劇場」，「作品」，「記念」のネットワークから劇場版の記念作品であること，「要素」を中心とするネットワークからミステリーやアクション，恋愛のストーリー要素があること，「赤い」を中心とするネットワーク

から赤色が運命を左右するラッキーカラーであることなどがわかる。

(2) 世紀末の魔術師

図4.2より，「怪盗」，「キッド」，「映画」，「登場」を中心とするネットワークから怪盗キッドが登場する映画であること，「要素」を中心とするネットワークから推理やアクション，ミステリー，歴史の要素があるストーリーであること，「ストーリー」，「面白い」のネットワークからストーリーが面白いことなどがわかる。

(3) 天国へのカウントダウン

図4.3より，「探偵」，「少年」のネットワークから，少年や探偵が活躍し，灰原と組織が強く関係する作品であることがわかる。実際，この作品では，主人公が所属する「少年探偵団」が活躍し，敵対組織である「黒の組織」に灰原が狙われるストーリーであり，このことがしっかり反映されていることがわかる。「ビル」，「ツインタワー」を中心とするネットワークから，作品の舞台の一つとして富士山の見えるビルであること，「シーン」のネットワークから，アクションが派手にあり観客の目を引くシーンであること，最後のシーンや脱出のシーンが最高であることがわかる。

(4) ハロウィンの花嫁

図4.4より，「警察」，「学校」，「刑事」のネットワークから警察学校の存在がストーリーに重要であることがわかる。実際，この作品では，「警察学校組」と呼ばれる存在が重要な存在である。「爆弾」を中心とするネットワークから首輪型の液体爆弾が事件に使われていること，「探偵」，「少年」，「ハロウィン」を中心とするネットワークからハロウィンを舞台として少年探偵団が活躍する作品であること，「シーン」，「アクション」を中心とするネットワークからアクションシーンが観客の目を引くシーンであることがわかる。

４．４．対応分析による評価スコアと作品の様相の関係

各作品の対応分析の結果を図4.5～4.8に示す。レビュー文とスコアの整数値による対応分析である。なお，スコアをつけていないレビュー文は用いていない。また，「人名」も共起ネットワーク分析と同様に抜いている。

(1) 時計じかけの摩天楼

図4.5より，評価5の方向に「素晴らしい」，「最高」があることより，評価5は一般的な高評価であること，評価4は原点近くにあることから多くの人が評価4をつけていること，

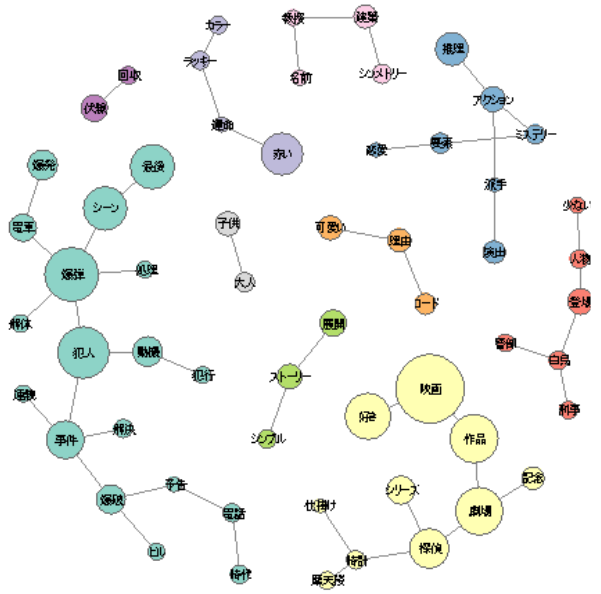

図4.1「時計じかけの摩天楼」の共起ネットワーク図

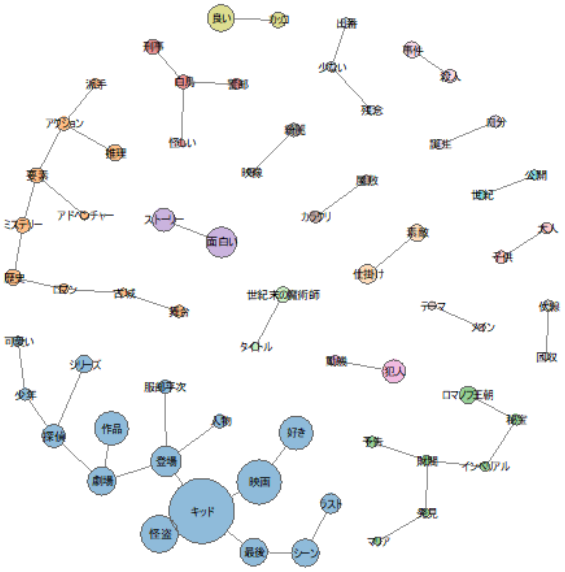

図4.2「世紀末の魔術師」の共起ネットワーク図

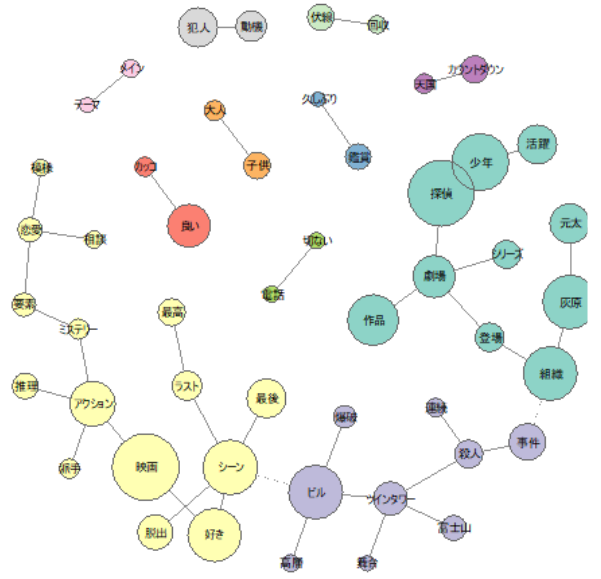

図4.3「天国へのカウントダウン」の共起ネットワーク図

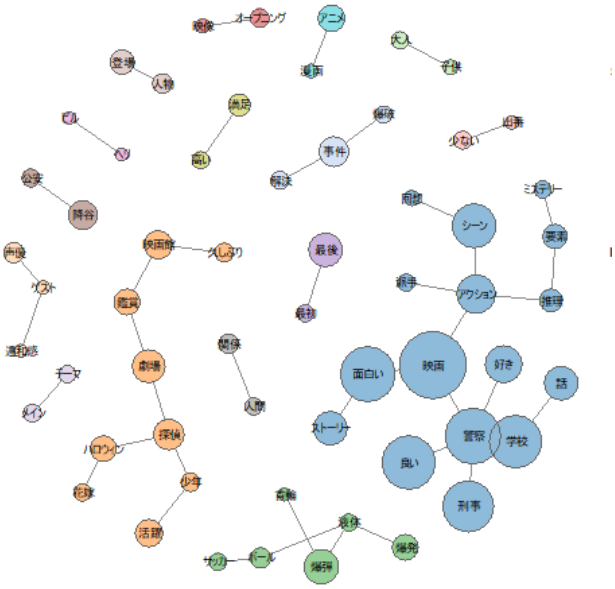

図4.4「ハロウィンの花嫁」の共起ネットワーク図

評価3の方向に「古い」があることより，作品の古さがマイナスの要因として作用していることがわかる。一方，評価2と評価1は全体の評価から大きく離れており，特異な評価となっている（実際のレビュー数の合計30件）ことがわかる。

（2）世紀末の魔術師

図4.6より，評価5の方向に「演出」，「主題歌」，「最高」，「美しい」があることより，評価5は演出や主題歌に高評価であること，評価4の方向に「要素」，「展開」，「面白い」，「楽しい」があることよりストーリーの内容や展開に高評価であること，評価3や評価2の方向に「話」，「難しい」，「出番」，「少ない」，があることより，評価3は話の難しさが，評価2はメインキャラクターの出番の少なさがマイナスの要因として作用していることがわかる。一方，評価1は全体の評価から大きく離れており，特異な評価となっている（実際のレビュー数4件）ことがわかる。

（3）天国へのカウントダウン

図4.7より，評価5の方向に「最高」，「クライマックス」，「大好き」があることより，評価5はクライマックスのシーンが高評価であること，評価4は原点近くにあることから多くの人が評価4をつけていること，評価3の方向に「爆破」，「印象」があることより，評価3は爆破に強く印象が残っていること，評価2や評価1は全体の評価から大きく離れており，特異な評価となっている（実際のレビュー数の合計20件）ことがわかる。

（4）ハロウィンの花嫁

図4.8より，評価5の方向に「感動」，「素晴らしい」，「最高」があることより，評価5は一般的な高評価であること，評価4の方向に「要素」，「アクション」，「爆発」「派手」があることより，評価4は爆発やアクションなどの派手な要素が高評価であること，評価3の方向に「残念」があることより，残念と判断される要素がマイナス要因として作用していること，評価2と評価1は全体の評価から大きく離れており，特異な評価となっている（実際のレビュー数の合計80件）ことがわかる。

５．結　論

本研究では，作者の文章作成の傾向や作品の内容など，対象の傾向や様相を文章データを基に明らかにすることで，情報縮約の機能が十分に働いているかを確認することを目的に，その典型的な事例として，ニュース時事能力検定の模擬問題サイトの全問題文と劇場版アニメシリーズの映画 4 作品の全レビュー文の計量テキスト分析を行った。

その結果，特徴分析と共起分析に加え，外部変数を用いた対応分析から，ニュース検定のデータでは，出題者全員が似たような語を用いているわけではなく，出題者によって出題する分野が異なり，傾向の差が明瞭に見て取れることがわかった。劇場版名探偵コナン

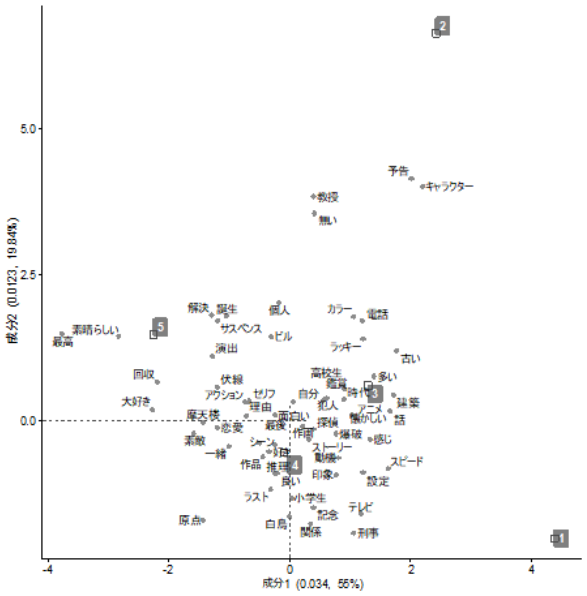

図4.5「時計じかけの摩天楼」の対応分析

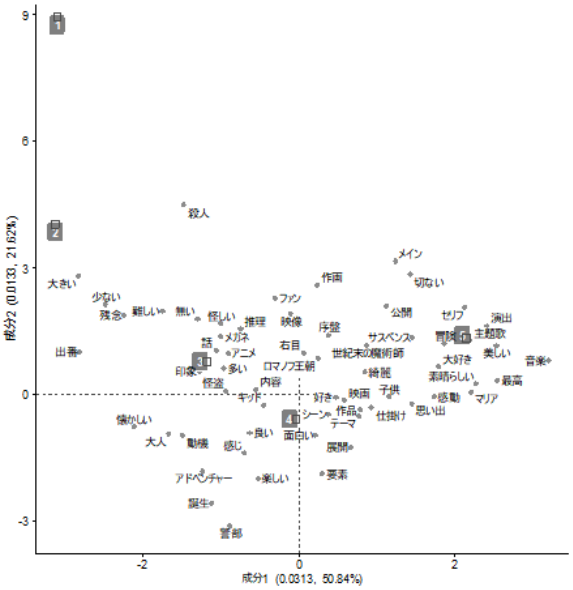

図4.6「世紀末の魔術師」の対応分析

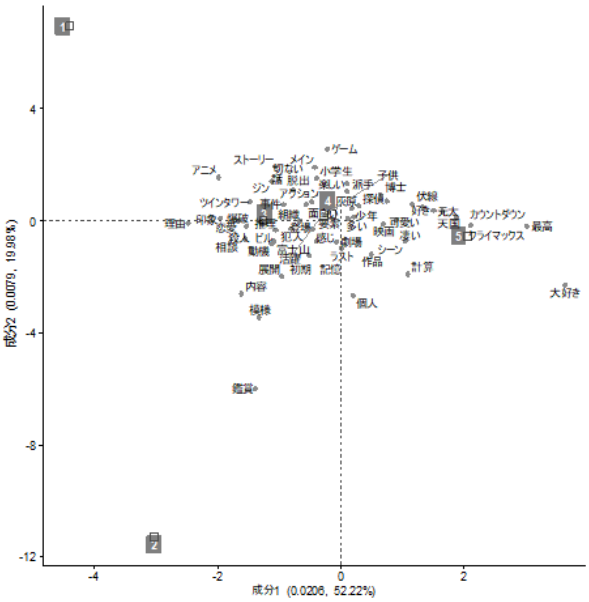

図4.7「天国へのカウントダウン」の対応分析

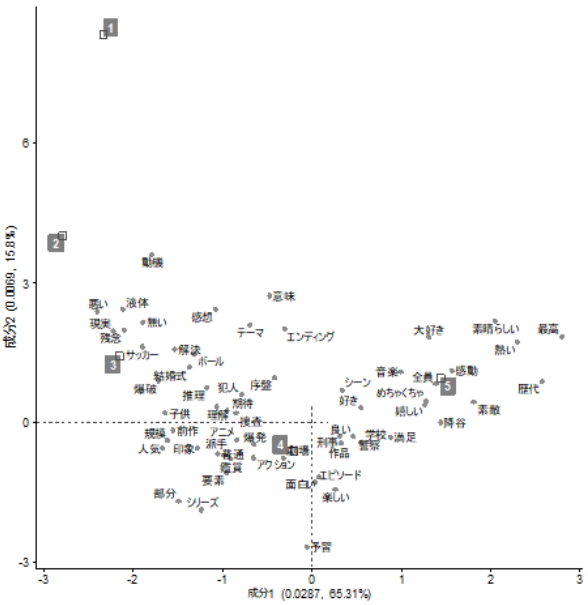

図4.8「ハロウィンの花嫁」の対応分析

のレビューデータでは，各作品のストーリーの主軸や要素，キーパーソンが明らかになり，実際に映画を鑑賞したときにわかるものと一致していることから，テキストマイニングにより情報縮約ができていることがわかった。

以上のように，検定試験の出題傾向および出題者の作問の特徴や，映画各作品の概要お

よび評価スコアとの関連を明らかにでき，判断や行動を起こすための情報提供ができることがわかり，テキストマイニングの情報縮約の機能が十分に働いていることが確認できた。

今後の課題として，固有表現を正確に抽出するために，オリジナル辞書を追加した形態素解析を行うこと，出題傾向の分析では，実際の過去問題の分析，出題年や時代背景との関係の分析，レビュー文の分析では，キャラクター性や感情の喚起の分析，投稿者の属性（性別や年齢）を加味した分析，複数のレビューサイトによる比較などがあげられる。

参考文献

上田太一郎・村田真樹・小木しのぶ・高山泰博・末吉正成・今村誠・渕上美喜（2008）．事例で学ぶテキストマイニング，共立出版.

岡山理科大学総合情報学部社会情報学科・経営学部．めざせニュース検定ニュース時事能力検定模擬問題，https://www.soci.ous.ac.jp/newskentei/index.html（2022.5.25）．

小川哲司（2022）．テキストマイニングとネットワーク分析を用いた映画評価の要因分析，経済経営論集，29（2），26-35.

金明哲（2009）．テキストデータの統計科学入門，岩波書店.

清水義孝（2022）．Python 最速データ収集術：スクレイピングでWeb情報を自動で集める，技術評論社.

清水義孝．図解！PythonでWEBスクレイピングを極めよう！，https://ai-inter1.com/ python-webscraping/（2023.1.19）．

株式会社つみき．Filmarks，https://filmarks.com/（2023.1.19）．

中島琢人・早田剛（2019）．計量テキスト分析を用いた柔道整復師国家試験問題の研究―柔道整復学に着目して―，環太平洋大学研究紀要，14，231-235

日本ニュース時事能力検定協会．ニュース時事能力検定とは，https://www.newskentei.jp/a_index.html（2023.1.19）．

樋口耕一・中村康則・周景龍（2022）．動かして学ぶ！はじめてのテキストマイニング，ナカニシヤ出版.

森巧尚（2019）．Python 2年生 スクレイピングのしくみ，翔泳社.