テキストマイニングにもとづくレビューのスコアリングを用いた映画の統計的分類

上智大学大学院理工学研究科数学領域・院　山縣　一慶

1　研究目的と背景

今日の日本は, 武士道に由来する武道や茶道などの伝統的様式のみならず, 映画や漫画, アニメなど現代文化が世界的に注目されており, それらは「クール・ジャパン」として総称される. 2010 年, 経済産業省製造産業局に” クール・ジャパン室” が設置され, 今では日本独自の産業や文化の世界進出を国が推進するまでになった. しかし統計手法を用いた現代文化の分析はまだ少なく, 今後国内外に日本特有の文化を発信するにあたり, 分析の需要は高まっていくと考えられる. 本研究では, 数ある” クールジャパン” の 1 つである映画に注目した. 映画は大衆文化として広く浸透しており, 視聴者によって投稿された口コミを閲覧できるサイトが大変多い. その一方, 映画の分類である” ジャンル” については明確な定義が存在しない. 我々が普段目にする映画のジャンルは, 主題や題材など映画を構成する諸要素の共通点をもとに, メディアが通俗的に分類したものである. そのため, 設定された映画ジャンルと映画の内容に差異を感じることも起き得る. 本研究では映画のジャンル分けに焦点を当て, テキストマイニングを用いてレビューから映画の特徴を表す語を抽出し, 抽出件数をもとにスコアリングすることで映画を統計的に分類することが目的である.

2　従来の研究手法と関連研究

2.1　テキストマイニングを用いた情報抽出の概要

テキストマイニングは, 文字列や自由記述されたテキストファイルに対するデータマイニングである. 単語や文節で区切り, それぞれの出現件数や出現傾向を解析することで有用な情報を取り出すことが目的である. 日本語で書かれたテキストデータを扱う場合, 文章の構成を品詞レベルで解析するために, 文章を形態素に分解する必要がある. 形態素とは, それ以上分割することができない語の最小単位である. 文章を読み込み, それぞれを形態素に分解する一連の工程を形態素解析という. 形態素解析を行うためのソフトウェアは無料で多々公開されており, 本研究では形態素解析ソフトの 1 つである TTM [6] を用いている. また, TTM では同義語を登録（以下, 辞書作り）することで, 同義語の出現件数を分析することも可能である. 下の表は, 辞書作りの一例である. 辞書 1 では, 我々が”SF” という語から連想するであろう語を登録したものである. それに対して辞書 2 は, インターネット上で公開されている同義語辞書（シソーラス）の Weblio [8] を用いて”SF” の同義語を登録した. また, 映画「スター・ウォーズ　フォースの覚醒」に関するレビュー 493 件に対して, 作成した辞書 2 つと TTM を用い, レビュー中で”SF” の語が使われている件数をレビュー点数ごとにまとめ, 比較した. 対象レビューは映画専門レビューサイトである Yhoo!映画 [7] から Ruby 言語で作成したプログラムを用いて抽出している.

辞書番号	単語	同義語
1	SF	宇宙, ロケット, 船, 星, 彗星, 惑星, 月, NASA
2	SF	エスエフ, サイエンス, フィクション, 空想科学小説, 近未来小説

辞書 1 を用いた抽出件数		
レビュー点数	抽出語	出現件数
1	SF	11
2	SF	16
3	SF	23
4	SF	24
5	SF	18

辞書 2 を用いた抽出件数		
レビュー点数	抽出語	出現件数
1	SF	3
2	SF	3
3	SF	1
4	SF	4
5	SF	5

辞書 1 は辞書 2 と比べ, SF の内容を表す語を多く含んでおり, SF に関する語を用いて書かれたレビューを多く抽出できていることがわかる.

2.2 統計手法を用いた単語の分類

テキストマイニングにおける辞書は語の出現件数に大きな影響を及ぼすため, 辞書の構築は慎重に行う必要がある. 従来の辞書作りでは, インターネット上の同義語辞書（シソーラス）を用いることが多い. 短時間で同義語を見つけることができるが, 分類対象の内容に合った同義語を抽出することができないという欠点を持つ. 統計手法を用いた同義語抽出の先行研究としては, 藤村 [3] が挙げられる. 藤村の研究では, ノート PC に関するレビューデータをもとに, 肯定的な意味を持つ語と否定的な意味を持つ語を分別するためのスコアリングを提案しており, ノート PC についてのレビューから語の分類に成功した. また, Anne [1] は ”personalization” と”customization” という曖昧な定義を持つ 2 語が分類可能か否か研究している. Anne は ”personalization” について書かれたテキスト 883 件, ”customizaiton” について書かれたテキスト 1544 件を対象に, テキスト内でそれぞれの語と同時に出現する語（以下, 共起語）の出現件数を調べた. そして, テキスト内である語と同時に出現する共起語の出現件数の相関から 2 語が分類可能であることを述べている. 本研究では, 共起語の出現件数（共起件数）から語が分類可能であることに注目し, 映画について述べられたテキストデータをもとに, 映画分類に適した辞書を作成した.

3 クラスター分析を用いた共起語辞書の作成

3.1 概要

本研究では, 映画レビューが, 映画の内容についての文章と, レビュー著者の感想について書かれた文章で構成されていると考えた. レビューにおいて映画内容について言及する際に使われる単語群をジャンル属性, 著者の感想について言及される際に使われる単語群を感情属性とした. シソーラスを用いて作成する辞書はジャンル属性として 22 語を設定し, 感情属性としては 8 語設定した. これら 30 語の同義語をインターネットで無料公開されているシソーラス, Weblio を用いて登録した辞書の一例を挙げる.

属性番号	単語	同義語
1	SF	エスエフ, サイエンス, フィクション, 空想科学小説, 近未来小説
2	ロマンス	恋愛, 恋愛関係, 恋路
3	西部劇	該当同義語なし

シソーラスを用いた辞書作りでは, ジャンル属性を抽出するための映画内容に関する語が少なく, 映画の分類に不適であることが推測できる. この問題を解決するために, 本研究では映画レビューサイトである映画.com[2] から映画 17000 本のあらすじを抽出し, 共起語の出現件数をもとに, ジャンル属性を再構築した.

3.2 作成手順

大型レビューサイトである映画.com には, 映画に対するレビューのみならず, 映画のあらすじ等様々な情報が公開されている. 本サイトから映画のあらすじについて書かれたテキストを, Ruby 言語で作成したプログラムを用い

て 17000 件抽出した. この 17000 件の映画あらすじに対して, 形態要素解析を行うことで名詞 3000 語を抽出し, 共起件数を求めた. 表は出現件数上位 10 個の名詞における共起件数の例である. 得られた 3000 × 3000 行列全て

	監督	撮影	彼	脚本	彼女	担当	音楽	それ	製作	出演
監督	13629	11079	8089	7939	6059	5942	5642	5396	5485	5351
撮影	11079	11876	7369	7333	5542	5953	5647	4928	5223	4936
彼	8089	7369	9533	5118	4846	4350	4451	3888	4147	4044
脚本	7939	7333	5118	8409	3968	4153	4113	3275	4061	3763
彼女	6059	5542	4846	3968	7127	3267	3219	3018	2991	3011
担当	5942	5953	4350	4153	3267	6487	4071	2727	3886	3670
音楽	5642	5647	4451	4113	3219	4071	6170	2350	4597	4115
それ	5396	4928	3888	3275	3018	2727	2350	6131	2196	2335
製作	5485	5223	4147	4061	2991	3886	4597	2196	5984	3918
出演	5351	4936	4044	3763	3011	3670	4115	2335	3918	5968

の要素に対して逆数をとり, 新たに得られた行列を語の類似度を表す距離行列とした, なお, 共起件数が 0 の場合, 定数 (0.01) を代入して逆数をとった. 共起件数から測定した 2 語の距離が長ければ, 語の関連性は弱く, 短ければ強いと推定できる. 得られた距離行列に対し, 実務的にも最も多く採用されているウォード法を用いてクラスター分析を行った. その後, クラスターを 43 個に分割し, 映画内容を表す特徴的な語が含まれる 12 のクラスターを抽出した. クラスター内の語は互いに関連性が強く, 共起語として扱うことにし, これら 12 語抽出すべき語として辞書に登録した. この共起語抽出によって得られた辞書を共起語辞書とする. ジャンル属性に含まれる語の共起語の一例である.

属性番号	単語	同義語
1	SF	銃撃戦, 開発, 地上, 変身, アジト, 黒幕, 地下, 爆弾, 標的, 暴走, 通報, 入手
2	ロマンス	日常, 青春ドラマ, 青春映画, ボーイフレンド, ガールフレンド, 仲直り, 魅了
3	西部劇	メキシコ, 荒野, 西部, アメリカ人, インディアン, 牧場, 保安官, 牛, 農場

4 共起語の出現件数を用いたスコアリング

本研究では, 大型映画データベースサイトである IMDB [4] の視聴ランキングに掲載されている映画の中でも, レビュー数が充実している映画 32 本を分類対象とした. 対象映画のレビューは, 映画専門レビューサイトである Yhoo!映画から, Ruby 言語を用いて作成したプログラムで抽出した. また, 本研究で提案するスコアリングは, 国土交通省 [5] で提案された公共事業の評価アンケートに対する重みづけの手法に, 対象データの語数による重みを加えたものである.

以下, 共起語辞書を用いたときのスコアリングについて説明する. 映画番号 i について書かれた, 評価点数 j のレビューにおける, ジャンル属性に含まれる単語の総出現件数を e_{ij} とし, 感情属性に含まれる単語の総出現件数を c_{ij} とする. また, ジャンル属性, 感情属性に含まれる特徴的な語の番号 (前節の図中, 属性番号と同一) を k とする. この時, 映画番号 i について書かれた評価点数 j のレビューにおける, ジャンル属性に属する属性番号 k の語の出現件数を c_{ijk}, 感情属性に属する属性番号 k の語の出現件数を e_{ijk} とする. 以上を用いて, スコアリングの定義を行った. 映画番号 i について書かれた評価点数 j のレビューにおけるジャンル属性と感情属性の重みを, それぞれ,

$$wc_{ij} = 100 * \frac{c_{ij}}{e_{ij} + c_{ij}}$$

$$we_{ij} = 100 * \frac{e_{ij}}{e_{ij} + c_{ij}}$$

とする. また, 映画番号 i における点数 j のレビュー総単語数を s_{ij} とする. このとき, 映画番号 i における, ジャ

ンル属性に所属する k 番目の特性が持つスコア C_{ik} を

$$C_{ik} = \sum_{j=1}^{5} \left(wc_{ij} \cdot \frac{c_{ijk}}{\sum_{k=1}^{22} c_{ijk}} \cdot \frac{s_{ij}}{\sum_{j=1}^{5} s_{ij}} \right)$$

とし, 感情属性に所属する k 番目の特性が持つスコア E_{ij} を

$$E_{ik} = \sum_{j=1}^{5} \left(we_{ij} \cdot \frac{e_{ijk}}{\sum_{k=1}^{8} c_{ijk}} \cdot \frac{s_{ij}}{\sum_{j=1}^{5} s_{ij}} \right)$$

とする. これらのスコアリングを用いて 32 本の映画すべてに対し各属性番号ごとのスコアをつけた.

5　スコアリング結果の統計的分析

シソーラスを基に作成した同義語辞書と, クラスター分析を基に作成した共起語辞書を用いて, スコアリングを行った. また, 両スコアリング結果に主成分分析を適用することで, 映画レビューから得られる各映画の傾向を可視化し, 比較を行った. 数値実験の結果はシンポジウム当日に発表する.

参考文献

[1] Anne, S. and Johanna, B., "Applying text-mining to personalization and customization", Expert Systems with Applications, 39, 10049 - 10058 (2012).

[2] 映画.com, http://eiga.com/

[3] 藤村 滋, 豊田 正史, 喜連川 優, Web からの評判および評価表現抽出に関する-考察", 社会法人情報処理学会研究報告, (2004).

[4] IMDB, http://www.imdb.com/chart/top/

[5] 国土交通省 公共事業評価システム研究会, " 公共事業評価の基本的考え方 (公共事業評価システム研究会報告) について", 国土交通省, (2002).

[6] TinyTextMiner β version , http://mtmr.jp/ttm/

[7] Yahoo!映画, http://movies.yahoo.co.jp/

[8] Weblio 類語辞典, http://thesaurus.weblio.jp/