ユーザの視点を利用した映画の分類

阿部 倫子† 田中 久美子‡ 中川 裕志††

†東京大学大学院学際情報学府 ,‡東京大学大学院情報学環, ‡†東京大学情報基盤センター

†abebe@r.dl.itc.u-tokyo.ac.jp, ‡kumiko@ipl.t.u-tokyo.ac.jp ‡†nakagawa@dl.itc.u-tokyo.ac.jp

1. はじめに

現在, パソコンやインターネットの普及, 通信回線の高速化に伴い, テキストのみならず, 音声や画像などのマルチメディアデータに関しても, より高度な検索方式が求められている. 例えば, 画像や映像の検索では, 画像の解析結果を用いる方法が多数研究されている[1]. また, 映像の音声トラックを音声認識した結果を使用する方法も研究されている[2].

また, インターネットは個人からの情報発信を容易にし, それによって, 商品の評判などの新たなメタな情報を生みだしている. このような情報は極めて人間の主観に依存する情報である. そこには, 多数の個人ユーザの観点や主観という情報が含まれている.

そこで, 本研究では映画という, 本来は非言語情報である対象を, コメントという言語情報を用いて分類することを試みた. コメントとは, 映画に対する個々のユーザの印象や感想, 評価が「言葉」で表されたものである.

本研究では, まずコメントを用いて, 機械学習によってどの程度の精度で映画の自動分類が可能かを探った. この結果を踏まえ, 次に人手で分類された映画をコメントの機械処理によって「再分類」を行った. 「再分類」とはやや分かりにくい概念であるが, ここでは, 人手で既に分類されている映画をコメントの機械処理によってもう一回分類し直す, という意味で「再分類」と呼んでいる. 映画に与えられたコメントには, ユーザの感想や評価などが表現されている. したがって, コメントを用いて分類を行うことによって「ユーザの視点や主観」が分類に取り入れられる. この再分類によって, データベースの作成者の主観に偏った基準ではなく, 多くのユーザの判断基準を反映させた分類を取り出すことができる可能性がある.

1.1. 映画評価サイト CinemaScape

本研究では, 映画評価サイト, CinemaScape[3]で収集されているコメント情報を利用した. CinemaScape は, 現在, 映画のコメント情報を集めた日本語のサイトとしては最大のサイトである.

CinemaScape に集められている, ユーザからのコメントの一例を(表 1)に示す. ここでは, 監督や出演者といった, 収録されている映画に関する基本的な索引情報はすべて Internet Movie.Database(IMDb)という映画データベースに基づいている. 映画は, あらかじめ18のジャンル(表 2)に分類されているが, この分類もIMDb による分類に準拠したものである. CinemaScape のデータの概要を(表 3)に示す.

表 1 コメントの一例

★5 親父はこの映画の大ファン. なので, ワケもわからなかった子どもの頃から, 無心の切り出し口上は「ゴッドファーザー, お願いがあるのですが…」だった.
★5 「ファミリー」に二重の含みがあるように, 「血」という言葉にも大切な意味二つ. そして, そのうちの「ありきたりではない方」の血がないことには成立しない, この家族の歴史の悲哀. 激情. 虚しさ. 寂しさ. そしてイヤになるほど鮮烈な, 美.
★5 マイケルになりたかった大学生の頃・・・

※(「ゴッドファーザー(1972/米)」より一部抜粋)

表 2 IMDb におけるジャンル

Action	Adventure	Animation	Comedy
Crime	Drama	Fantasy	Horror
Musical	Mystery	Romance	SciFi
Thriller	War		

表 3 CinemaScape データの概要

映画総数	9413
コメント有り映画数	7004
コメント投稿ユーザ数	1800
単語総数	1109877
異なり数	55175
平均コメント数(コメント/1映画)	12
平均単語数(単語/1映画)	158
平均ジャンル数(付与されているジャンルの数/1映画)	2.4

1.2. コメントの性質

コメントデータはユーザの自発的な行為として形成され, 日々刻々とWeb上に増え続け, 蓄積されていく. また, 再分類の目的のもとでは, コメントを用いるからこそ, ユーザの主観を反映させ, 意義ある情報を発掘することが期待できる.

しかし, 通常の文書分類と異なり, コメントは, 映画そのものではない. 映画のコメントは, まだその映画をみ

ていない読者を想定して書かれるものであり，必然的に抽象度の高い記述になる．よってそこには，ストーリーなど，具体的な内容については現れにくいと考えられる．こういったコメントの性質は，それぞれ利点，問題点として以下のようにまとめられる．

利点：ユーザの主観という付加的な情報を有する．よって機械分類を行うことにより，単独の分類作成者（この場合 IMDb）の判断基準を離れ，多くの個人ユーザの主観を捉えることができる．

問題点：ストーリーなど，映画の内容についての情報が不足しがちである．その際，ジャンルを特徴づける表現が同時に不足してしまえば，的確な分類の予測が困難なものになる．

なお，映画についての情報としては，言語で表現されたメタ情報としてはコメント以外にも監督名，出演者名，制作会社，また，あらすじや台本（特に台詞）のようなデータがある．これらのデータを用いることは，ジャンルへの分類には効果を発揮するかもしれない．しかし，この研究での目的が，個々のユーザの主観や好みに対応する分類であることを考えれば，これらの中立的情報の効果は多くを期待できない．したがって今回の実験では使用していない．

1.3. コメントの表現力

前節においてコメントの持つ問題点として述べたように，映画そのものの具体的内容がコメントに現れなければ，いかに優れた分類器を用いようとも，分類精度の向上を望めない．そこで，実際にコメントはどの程度，ジャンルに対して表現力を持つのかを探るため，アンケートによる分析を行った．

アンケートでは，まず，ある映画に与えられたコメントを，映画のタイトルとジャンルは伏せて被験者に提示する．被験者は提示されたコメントが，どのジャンルの映画に与えられているものなのかを推測し，表 2にあげた18 のジャンルから選択する．

コメントの表現力が十分であれば，人間にとっても，どのジャンルに当てはまるコメントなのかをより推定しやすいと考えられる．また，人間による分類を Gold-Standard と見れば，機械分類の結果に対する一つの指標になる．このアンケートの結果を示す．

表 4 人間によりコメントから分類を予測

回答者数	79
一人あたり検証映画数	28.2
再現率	0.37
適合率	0.48
F 値	0.42

この結果が示すように，人手によって分類を行った場合でも，コメントをみるだけでは，分類が難しいことがわかる．アンケートでは，コメントのジャンルに対する表現力の弱さを窺わせる結果となった．以降で，コメントを用いた分類の実際について述べる．

2. 分類モデル

分類器としては，本研究の目的である自動分類と再分類のどちらにも同じアルゴリズムを用いる．分類モデルの構築には，ナイーブ・ベイズを用いた．

ナイーブ・ベイズによる分類モデルによれば，映画 m の属するカテゴリー c を決定する問題は，事後確率 $P(c \mid m)$ を最大化するようなカテゴリー $\hat{c}$ を求める問題である．これを，

$$\hat{c} = \arg\max_{c_i} P(c_i \mid m) \tag{1}$$

と表す．このとき，IMDb には 18 のジャンルが存在するので，i は 1～18 の値をとる．ここで，映画 m はコメントに含まれる単語 w_k の集合によって表されるとすると，

$$\hat{c} = \arg\max_{c_i} P(c_i \mid w_1, \cdots, w_n) \tag{2}$$

ここでベイスの定理により

$$\hat{c} = \arg\max_{c_i} P(w_1, \cdots, w_n \mid c_i) P(c_i) \tag{3}$$

となる．さらに，単語の独立性を仮定し，

$$P(w_1, \cdots, w_n \mid c_i) = \prod_i P(w_n \mid c_i) \tag{4}$$

とおき，式(3)，に代入すると，

$$\hat{c} = \arg\max_{c_i} P(c_i) \prod_{k=1}^{n} P(w_k \mid c_i) \tag{5}$$

が導かれる．

また，c_i における w_k の出現回数を F_{ik}，c_i に出現する単語の総数（のべ数）を N_i とすると，$P(w_k \mid c_i)$ は，

$$P(w_k \mid c_i) = \frac{F_{ik}}{N_i} \tag{6}$$

$P(c_i)$ は，

$$P(c_i) = \frac{c_i \text{の映画数}}{\text{全映画数}\ (=7004)} \tag{7}$$

と定義される．

式(6)において，単語によっては $F_{ik} = 0$ となるので，式(5)の右辺が 0 となってしまう．これを避けるため，予期尤度推定法[4]によりディスカウンティングを行った．予期尤度推定法では単語の頻度にあらかじめ 0.5 を足しておく方法である．単語の異なり数を V_{all} とおくと，ここでは，

$$P(w_k \mid c_i) = \frac{F_{ik} + 0.5}{N_i + 0.5 * V_{all}} \tag{8}$$

となる．

なお，コメントから単語を切り出す際には，JUMAN Ver.3.61 により形態素解析を行い，助詞，助動詞，接

続詞以外の単語を用いることとした.

3. 自動分類

ナイーブ・ベイズによる分類では,結果は順位付き(1位のジャンル～18位のジャンル)で出力される.よって,評価には2つの方法を用いた.

まず一つ目は,通常の再現率や適合率を用いるものである.このためには,まず,ナイーブ・ベイズの順位付きの結果に対して,なんらかの閾値によってジャンルを決定しておく必要がある.本研究では,各映画において,IMDbによりあらかじめ付与されていたジャンルの数をxとした場合,ナイーブ・ベイズの出力に対する上位x位までをその映画のジャンルとした.

再現率,及び適合率は,IMDbによって付与されていたジャンルを正解集合と見なし,各映画において,IMDbによって付与されていたジャンルの数をG_{imdb},ナイーブ・ベイズにより付与したジャンルの数をG_{bayes},ナイーブ・ベイズが付与したジャンルのうち正解であったジャンルの数を$G_{correct}$ ($G_{imdb} \cap G_{bayes}$),とすると,

再現率=$G_{correct} / G_{imdb}$ (9)

適合率=$G_{correct} / G_{bayes}$ (10)

定義される.

表 5 平均適合率の計算法

シックス・センス(1999／米)

IMDbによるジャンル	Thriller, Drama, Horror	
ナイーブ・ベイズによるジャンルの順位	1 Drama	← 1/1
	2 Thriller	← 2/2
	3 Comedy	
	4 SciFi	
	5 Action	
	6 Romance	
	7 Crime	
	8 Mystery	
	9 Horror	← 3/9
	10 Adventure	
	11 :	
	(1/2 + 2/2 + 3/9) / 3 ≒ 0.778	

二つ目に,平均適合率(Average precision)[5]を用いた評価を行った.平均適合率を用いることで,順位付き分類結果そのままの状態を考慮し,また,再現率と適合率を総合的な観点から1つの値で評価することができる.表 5に示すように,平均適合率は,各映画におけるナイーブ・ベイズによるジャンルの順位に対し,IMDbで付与されているジャンルが出現したそれぞれの時点での適合率を計算し,それらの適合率を平均したものである.上位X位での適合率を

適合率(X)=Xより上位の正解数／X (11)

とし,分類結果のうち正解であるジャンルの順位を表す集合をSとおくと,

平均適合率 = $\text{Average}_{(N \in S)}${適合率(X)} (12)

となる.

これらすべての評価実験において,データを訓練集合とテスト集合に9:1の割合で分割し10-fold交差検定を行った.その結果を表 6 に掲げる(右列は,訓練集合そのものを分類した結果).

表 6 ナイーブベイズ法による自動分類の評価結果

	10-fold 交差検定 (自動分類)	訓練集合=テスト集合 (再分類)
平均適合率	0.70	0.92
再現率=適合率(=F値)	0.58	0.87

この結果が示すように,自動分類としてはさほどよい結果は得られていない.これは,コメントという情報源の性質の問題であるといえよう.つまり,先に述べたように,コメントは映画そのものの具体的な内容記述を表しておらず,その結果,自動的な分類を困難にしていると考えられる.だが一方で,人間による分類の推測においてF値が0.42であることを鑑みると(表 4),機械分類によってもすでに最善が尽くされていると考えられる.よって自動分類に関しては,コメントを用いたアプローチの限界に近い結果を示していえそうである.

4. 再分類

次にコメントの持つもう一つの側面である「ユーザの主観を含む」という特徴を生かすための,コメントを用いた機械学習による再分類の試みについて述べる.用いる機械学習の方法は,自動分類と同じナイーブベイス分類である.

未知データに対するカテゴリーを予測するという自動分類と異なり,再分類の目的は,多数の個別ユーザの主観を反映させた新たな分類をとりだすということである.この再分類の目的のもとでは,利用できる最大限の情報を用いてエラーを回避し,コメントの表現力の弱さを補えばよい.つまり,テスト集合と訓練集合を区別せず,全データを用いて分類すればよい.

実際に訓練集合そのものを分類すると,利用できるすべての情報を用いて,分類の学習を行っているにもかかわらず,IMDbの分類と異なる分類を返す映画が出現する.表 6によれば,ナイーブ・ベイズ分類では,1割程度のずれを生んでいることが分かる.このような映画の一例を表 7に示す.この「分類のずれ」が,再分類のもたらした「新しい情報」である.さらにこの「分類のずれ」は下の3つに類型化できる(ただしこれらは明確に区別できるものではない).

1) 個々のユーザの主観を反映させた意外性を譜含むが間違いではない,という新たな分類
2) IMDbによる分類を補完する分類
3) 機械分類の誤り

1)は，ユーザならではの主観が反映されたものである．たとえば，映画『ドラえもん　のび太の創世日記(1995/日)』がSciFiに分類される場合である．2)は，IMDbの分類作成者の誤りや見落としを補完する類のものである．

表 7 分類のずれた映画

映画	IMDb	ベイズ1位
ダーティハリー4	Crime Drama	Action
ガメラ対宇宙怪獣バイラス	Drama	Action
バンディッツ	Drama	Comedy
タワーリング・インフェルノ	Drama	Action
仕立て屋の恋	Thriller Crime	Drama
キャスパー（1995／米）	Adventure	Comedy
シャーロックホームズの冒険	Drama	Mystery
ドラえもん　のび太の創世日記	Animation	SciFi

このうち再分類において，取り出したい情報は，1)と2)である．そこで，「分類のずれ」に対し，アンケートによる評価実験を行った．アンケートは，「ナイーブ・ベイズによって1位に出力されたジャンルが，IMDbによる分類中には含まれていない」という676件の映画を「分類のずれ」を生じている映画とし，これらに対し，ナイーブ・ベイズの付与したジャンルが「ふさわしいか／ふさわしくないか」を判断してもらうというものである．

このアンケートでは，たとえ元のIMDbのジャンルとは異なっていても，ナイーブベイズ分類の付与したジャンルを「ふさわしい」とする回答が 52%にのぼった(回答者89人・1人あたり平均29件の映画について回答)．

表 8 分類のずれを評価

回答者数	89
総回答数	2588
ふさわしい	51.7%
ふさわしくない	48.3%

さらに，この「ふさわしい」とされたもののうち，類型1)なのか2)なのかを問うアンケートを行ったところ，前者が29%，後者が71%という結果を得た．あらすじなどのコメント以外の何らかの言語情報を用いて映画を分類した場合でも，類型2)の情報は得られると考えられる．しかし類型 1)の分類は，コメントを用いるからこそ得られる情報である．したがって，個々のユーザの主観を反映した分類への足掛かりを得たといえる．

表 9 分類のずれにおける類型

回答者数	11
総回答数	335
類型 1)	28.9%
類型 2)	71.1%

5. まとめ

コメントを用い，自動分類を行った場合には，ジャンルに対する表現力が弱いというコメントの性質が分類に悪影響を及ぼすと考えられ，機械学習による分類でも人手による分類精度の約 40%を上回りはしたものの，実用的といえるレベルまでは向上しなかった．

しかし，ユーザからの評価情報という特殊な情報を含むものとしてコメントをとらえた場合，再分類を行うことによって，ユーザの主観が反映された分類という新しい情報を取り出せることが分かった．この再分類は，IMDb において与えられていた単一の分類作成者の視点から，多くの個別ユーザの主観へと，分類基準を移動させているといえる．コメントを用いるからこそ，このような意義のある再分類が可能となる．

従来の文書分類に関する研究では，分類そのものを自動化すること主たる目的としていたため，そこでは，分類モデルの精度を競うことのみに注力する傾向が強い．そのような中で，今回の再分類の実験は，機械分類の新しい可能性を見いだしたものであり，今後の応用や対象とするデータの拡大も期待できると考える

コメントを用いた実用的，汎用的な再分類システムを構築するためには，今後の課題として，「分類のずれ」のうち，有用な分類のずれ(類型1)，2))と，無用な分類のずれ(類型 3))を機械的に判別可能にすることが必要である．これは，「分類のずれ」に対するユーザの評価となんらかの情報の間に関連性を発見することができれば，解決可能であると考えられる．

文　献

[1] Henrich, A., The LSDh-Tree: An Access Structure for Feature Vectors. Proceedings of the 14th International Conference on Data Engineering (ICDE 1998). p.362-369 (1998).

[2 Satoh, Shin'ichi., Nakamura, Yuichi., Kanade, Takeo. Name-It:Naming and Detecting Faces in News Videos, IEEE MultiMedia, Vol.6, No.1, pp. 22-35 (1999).

[3] 舘村純一. CinemaScape. http://cinema.media.iis.u-tokyo.ac.jp/

[4] Good, I. J., The Estimation of Probabilities, MIT Press Cambridge, MA (1965).

[5] Schuetze, H., Manning, C., Foundations of Statistical Natural Language Processing. MIT Press, Cambridge MA, p.534-536 (1999).