自然言語処理　150－16
(2002. 7. 16)

コメントを用いた映画の分類

阿部 倫子[1]　田中 久美子[2]　中川 裕志[3]

[1]東京大学大学院学際情報学府 〒113-0033 東京都文京区本郷 7-3-1

[2]東京大学大学院情報学環 〒113-8656 東京都文京区弥生 2-11-16

[3]東京大学情報基盤センター 〒113-0033 東京都文京区本郷 7-3-1

E-mail: [1]abebe@r.dl.itc.u-tokyo.ac.jp, [2]kumiko@ipl.t.u-tokyo.ac.jp [3]nakagawa@dl.itc.u-tokyo.ac.jp

あらまし　映画情報サイトに集められたユーザからの映画に対するコメントを用いて，ナイーブ・ベイズ分類により個々の映画を分類し，既存のジャンル分けと比較評価した．分類精度の客観的評価には平均適合率を用い，10回の実験において平均で約 0.7 程度の分類精度を示した．実験における個々の映画の分類を詳細に観察すると，既存の分類と機械による分類が異なっている場合にも，機械による分類情報が有用な情報をもっている場合があることに気づく．今後の課題としてこれらの情報をいかに映画の探索システムの中にいかしていくかがある．このための予備データとしてナイーブ・ベイズ分類が既存の分類とは異なるジャンルになった場合の例についても、その内容を分析した結果を報告する。

キーワード　自動分類，ナイーブ・ベイズ分類，映画探索システム

Classification of Films using Comments

Michiko ABE[1]　Kumiko TANAKA[2]　and　Hiroshi NAKAGAWA[3]

[1]Interfaculty Initiative in Information Studies Graduate School of the University of Tokyo　7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan

[2]Interfaculty Initiative in Information Studies Graduate School of the University of Tokyo　2-11-6 Yayoi, Bunkyo-ku, Tokyo, 113-8656 Japan

[3]Information Technology Center, the University of Tokyo　7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan

E-mail: [1]abebe@r.dl.itc.u-tokyo.ac.jp, [2]kumiko@ipl.t.u-tokyo.ac.jp [3]nakagawa@dl.itc.u-tokyo.ac.jp

Abstract　Using users' comments from a movie information site, movies were classified based on Naive Bayes classification. Evaluation was done bycomparing the results to existing genre classification. For the objective evaluation of the classification precision, we employed the average precision. We attested an average precision of about 0.7 over 10 experiments. When we look into the movie classification from this experiment detailedly, we notice that even in the case in which the existing genre classification and machine classification differ, the machine classification data holds valuable information. As future research, we want to investigate how to incorporate this data into a movie search system. As preparatory data in order to achieve this goal, we will report the results of our analysis of cases in which the Naive Bayes classification and existing classification differ.

Keyword　Naive Bayes classifier, Automatic classification, Movie search system

1. はじめに

現在，インターネットでは様々なデータベースが利用できる．また，ユーザからの情報発信，つまり評価や採点などを利用して，より充実した情報提供を行うサイトが数多くあり，情報発信・情報収集の場として多くのユーザを集めている．映画，本，レストラン，電気製品などその分野は多岐にわたる[1][2][3].

現状ではこれらのサイトは人手による管理に依存する部分が大きく，データ量も膨大である．より柔軟で多角的なデータベースの探索を可能にするには，ユーザから得られた情報を自動的に処理できる機能が望まれる．

本研究では映画情報サイト CinemaScape[4]に集められ

た，ユーザの映画に対するコメント情報を用いて，映画の分類を行った．映画そのものはテキストデータではないが，このようにユーザの映画に対するコメントに着目することで，映画の分類についても自然言語処理技術で扱いうるものとなる[5]．また，コメントの類似性を映画の類似性ととらえれば，新たな側面からの情報をユーザに提供することができると考える．

実験では分類手法としてナイーブ・ベイズ分類を用い，得られた結果を評価・考察した．ナイーブ・ベイズ分類は確率モデルに基づく分類法である[6]．また，分類結果を詳細に考察することで，人手で与えられた分類とナイーブ・ベイズ分類による自動分類で相違がある映画に関して，必ずしも自動的な分類が間違っているとはいえない映画が多々あることに気づく．

本研究では，ナイーブ・ベイズ分類によりコメントから映画をある程度分類することができることを実験により検証した上で，「分類」を単なる参考情報としてユーザに提示するだけではなく，映画をより多角的に探索できるシステムを考案することを目指している．

2. CinemaScape

インターネットで利用できる映画データベースにも様々なものがある．中でも最も知られているのは The Internet Movie Database (IMDb)[7]である．IMDb には 25 万件に及ぶ世界中の映画が収録されている．日本語で利用可能なデータベースには，allcinema ONLINE[8]，ぴあシネマクラブ[9]などがある．allcinema ONLINE は，サイトを訪れたユーザが映画に対するコメントを書き込むことができるのが特徴である．ぴあシネマクラブでは，収録している 17000 件の映画すべてについて，あらすじを得ることができる．

本研究で利用したコメント情報は，すべて CinemaScape で収集されているものである．allcinema ONLINE 同様，CinemaScape においても，コメント情報が収集されているのだが，allcinema ONLINE をはるかに上回るコメント情報が CinemaScape には蓄積されている．また CinemaScape ではコメントと同時に映画に対する 5 段階の採点情報を収集しており，これを用いて，協調フィルタリングによる映画推薦システムが実現されている[10]．

CinemaScape に収録されている映画に関する基本情報は前述の IMDb が使用されている．映画は 18 のジャンルに分類されており，この分類も IMDb による人手の分類に準拠したものである．（複数の分類が付与されている映画もある．）

登録されている映画数は 9413 件（2002 年 1 月現在）であるが，このうちどのジャンルも付与されていない映画と，ユーザからコメントがひとつも得られていない映画は，本研究では処理の対象からはずした．これにより，実際に本研究で使用した映画の総数は 7003 件となっている．コメントは，映画ごとに形態素解析をし，名詞，動詞，形容詞，形容動詞，未定義語，副詞，連体詞，感動詞をとりだした．また，半角文字はすべて全角に，アルファベット大文字はすべて小文字に変換した．

各映画において，コメントから得られる単語の数にはばらつきがある．有名な作品には多くのコメントが寄せられるが，あまり知られていない映画はコメントも少ない．1 映画あたりの単語数を平均すると約 10 語程度であり（最頻も 10 語，最大は 3797 語），通常の文書分類にくらべ比較的短いテキストをもとにしているので，分類が難しい場合もある．コメントデータの概要を（表 1）に示す．

（表 1） コメントデータの概要

	映画数	コメント数	単語総数	異なり数
Total	7004	111301	1109877	55175
Action	1443	28742	293145	27827
Adventure	439	9929	101459	15487
Animation	297	5652	57277	10739
Comedy	1774	28018	280043	27690
Crime	572	10120	102454	15847
Documentary	108	750	7982	3272
Drama	3442	53751	565513	39683
Family	136	1471	15282	4719
Fantasy	206	4076	42912	9168
Horror	440	7560	76203	13053
Musical	188	2483	26513	6928
Mystery	393	6503	66604	12602
Romance	902	16078	165460	20390
SciFi	561	14483	146445	18987
Short	79	552	4885	2312
Thriller	860	21072	212217	23013
War	271	5764	61723	11594
Western	148	1410	14229	4655

（図 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／Action・Crime・Drama）」より一部抜粋）

3. ナイーブ・ベイズ分類

3.1. ナイーブ・ベイズ分類の適用

ナイーブ・ベイズ分類は文書の分類法として，広く知られた方法である．文書分類についてはほかにも Support

Vector Machine[11]によるもの，決定木による分類[12]など，様々な方法が提案されている．しかし，本稿の目的は分類精度をみることよりも，コメントを用いて映画や商品を扱いうるのかどうか，その可能性を探求することを第一の目的としている．したがって，分類結果が分類方法になるべく依存せず，また，分類後の解析が簡単に行える単純なモデルを用いたい．このため，ナイーブ・ベイズ分類を選んだ．

実験では，1件の映画を1件の文書ととらえ，映画に対するコメント中に含まれる単語を，その映画（文書）に含まれる単語として扱った．

各ジャンルを $\{c_i;c_1,\ldots,c_{18}\}$，各映画を $\{m_j;m_1,\ldots,m_{7004}\}$ とおき，m_j に与えられているコメントにあらわれる単語を $\{w_k;w_1,\cdots,w_n\}$ とおくと，m_j に対してのジャンルは，事後確率 $P(c_j|m_i)$ を最大化するようなカテゴリ $\hat{c}$ 以下の式で求めることができる．

$$\begin{aligned}\hat{c} &= \arg\max_{c_i} P(c_i|m_j)\\ &= \arg\max_{c_i} P(c_i|w_1,\cdots,w_n)\\ &= \arg\max_{c_i} P(w_1,\cdots,w_n|c_i)P(c_i) \quad \cdots(1)\end{aligned}$$

さらに，各ジャンルのもとで単語は独立に生起すると仮定し，

$$P(w_1,\cdots,w_n|c_i)=\prod_{k=1}^{n}P(w_k|c_i)$$

とする．これにより，映画の分類は次式により行うことができる．

$$\hat{c}=\arg\max_{c_i} P(c_i)\prod_{k=1}^{n}P(w_k|c_i) \quad \cdots(2)$$

ここでは，

$P(c_i)=c_i$ に含まれる映画数 / 全映画数 $\cdots(3)$

とし，また，c_i に出現する単語総数を N_i，c_i において w_k が出現する回数を F_{ik} とおくと，

$$P(w_k|c_i)=F_{ik}/N_i \quad \cdots(4)$$

と定義する．

上記のように，個々の映画においてすべてのジャンルに対し，その事後確率をもとめることで，各映画に対する適切なジャンルを順位付けることができる．

3.2. ゼロ頻度問題

ところで，式(4)において，単語によっては，ジャンル c_i において w_k が出現する回数 F_{ik} が 0 となる場合がある．この場合，$P(w_k|c_i)=0$ となり，出現回数 0 の単語がひとつでもあれば，そのジャンルの事後確率は 0 という結果になってしまう．これを避けるためには，単語の出現回数の補正（ディスカウンティング）を行う必要がある．

ディスカウンティングには予期尤度推定法（ジェフリース・パークス法）[13]を採用した．予期尤度推定法は単語の頻度に 0.5 をあらかじめ足しておく方法で，すべての映画についての単語の異なり総数を V_{all} とおくと，$P(w_k|c_i)$ は以下の式で表される．

$$P(w_k|c_i)=(F_{ik}+0.5)/(N_i+0.5V_{all}) \quad \cdots(5)$$

ここで V_{all} は，単語の出現確率の合計が 1 になるように導入された定数である．各ジャンルにおいて一度も出現しない単語（0 頻度）の出現確率は

$$P(w_k|c_i)=0.5/(N_i+0.5V_{all}) \quad \cdots(6)$$

として得られる．

たとえば「Horror」というジャンルの中に，単語{A, B, C, D, E}が出現するとする．これらの単語の「Horror」における出現確率 $P(w_k|c_{horror})$ をそれぞれ{a, b, c, d, e}とする．さらに「God Father」という映画の中に，単語{A, B, D, F}が 1 回ずつ出現するとすれば，ジャンル「Horror」の「God Father」に対するベイズ事後確率は以下の式で求められる．

$$\begin{aligned}&P(c_{horror}|God\ Father)=\\ &\quad a*b*d*\{0.5/(N_{horror}+0.5*V_{all})\} \quad \cdots(7)\end{aligned}$$

4. 実験と評価

まず，7004 件の映画を，ランダムに 10 等分し，9:1 の訓練集合とテスト集合のペアをつくった（test1～10）．さらに，訓練集合に含まれる映画で分類の学習をおこない，テスト集合に含まれる映画（それぞれ約 70 件）を分類する実験をそれぞれのペアで行う 10-fold 交差検定を行った．

評価の尺度には平均適合率（Average precision）を用いた[14]．平均適合率を用いることで，順位付き分類結果を考慮し，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 4）ナイーブ・ベイズ分類による各ジャンルの順位
（例：シックス・センス）

シックス・センス（1999／米）	
IMDb による分類	Thriller／Drama／Horror
ナイーブ・ベイズ分類による順位	1 Drama 2 Thriller 3 Comedy 4 SciFi 5 Action 6 Romance 7 Crime 8 Mystery 9 Horror 10 Adventure 11 War 12 Fantasy 13 Animation 14 Musical 15 :

具体的には，各映画におけるナイーブ・ベイズ分類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精度を計算し，それらの精度を平均したものが平均適合率になる．

（表 4）にあらわした，映画「シックス・センス」の分類結果の場合，精度は 1 位 Drama の時点で 1/1，2 位 Thriller の時点で 2/2，9 位 Horror の時点で 3/9，となり平均適合率は 1/2 + 2/2 + 3/9 ≒ 0.778 となる．

10 回の実験それぞれにおける平均適合率を（表 5）にしめす．

（表 5） 10 回の実験における平均適合率

test1	0.706
test2	0.710
test3	0.698
test4	0.697
test5	0.709
test6	0.721
test7	0.696
test8	0.694
test9	0.690
test10	0.692
平均	0.701

5. 考察

IMDb により付与されているジャンルを正解集合とし，平均適合率を用いて分類精度の評価を行った．これにより，コメントを利用して映画をある程度自動的に分類できることがわかった．しかし，本研究が最終的に目指しているのは，このような評価尺度に基づいた映画の分類精度をあげることではない．人手による分類は，あくまで「どこかでだれかが」とりきめた分類基準に基づいており，そこに主観性が入ることは否めない．その分類を唯一無二の正解とし，評価を行ったところで，ユーザにとって本当に有意なものなのかを測ることにはできない．

本研究が，映画の「あらすじ」や「せりふ」ではなく，「ユーザからのコメント」を処理の対象として用いたのは，“コメントを用いることで，より多くのユーザからの意見を反映した「再分類」が行えるのでは”，という仮定による．計算機は人々のコメントから素直にジャンルを予想したにすぎない．人手による分類と機械による分類が異なる結果を示していても，それは単に，IMDb における分類基準とユーザの意見が異なっているからであるともいえる．よって，多くのユーザによって，「主観的な評価・感想」として集められたコメントに基づいて分類を行えば，ユーザにとってはより有意義な情報を提供できる場合もあると考える．人手による分類と，機械による分類の差にこそ意義があるといえよう．

そこで，IMDb による分類と，ナイーブ・ベイズ分類の結果が異なっている映画に関して，具体的に観察・分析を行った．ナイーブ・ベイズ分類により，1 位に順位付けられたジャンルにもかかわらず，IMDb ではそのジャンルに分類されていなかった映画の例が（表 6）である．ここでは極端な例を示すため，「訓練集合（7004 件の映画を含む）＝テスト集合」として分類を行った場合の結果を示す．

（表 6）にあげた映画の中でも，ナイーブ・ベイズ分類が 1 位にあげているジャンルが，あながち間違ってはいない印象

（表 6） IMDb とベイズ分類で結果が異なる映画の例

タイトル	IMDb による分類	ベイズ分類 1 位	2 位	3 位
タワーリング・インフェルノ	Drama	Action	Drama	Thriller
ダーティハリー4	Crime/Drama	Action	Crime	Drama
ガメラ対宇宙怪獣バイラス	Drama	Action	SciFi	Drama
空軍大戦略	War	Action	War	Drama
うる星やつら いつだってマイ・ダーリン	Animation	Action	Comedy	Animation
超音ジェット機	Drama	Adventure	Drama	War
サーキットの狼	Action	Animation	Action	Adventure
ルパン三世 念力珍作戦	Comedy	Animation	Action	Comedy
プラン9・フロム・アウタースペース	SciFi/Horror	Comedy	SciFi	Drama
現金に手を出すな	Thriller	Crime	Thriller	Romance
ハバナ	Drama	Crime	Thriller	Drama
山口組三代目	Action/Drama	Crime	Action	Drama
ブエナ・ビスタ・ソシアル・クラブ	Documentary	Drama	Romance	Comedy
あの夏，いちばん静かな海．	Romance	Drama	Romance	Crime
仕立て屋の恋	Thriller/Crime	Drama	Romance	Crime
ときめきメモリアル	Drama/Romance	Horror	Animation	Thriller
小人の饗宴	Drama	Musical	Drama	Animation
シャーロックホームズの冒険	Drama	Mystery	Adventure	Action
イン・ベッド・ウィズ・マドンナ	Documentary	Romance	Drama	Comedy
小さな兵隊	War	Romance	Drama	Comedy
パリの恋人	Comedy/Musical	Romance	Comedy	Musical
ネバーエンディング・ストーリー3	Fantasy	SciFi	Comedy	Action
ドラえもん のび太の創世日記	Animation	SciFi	Drama	Animation
アルカトラズからの脱出	Drama	Thriller	Drama	Action
戦略空軍命令	Drama	War	Drama	Action
ジャンヌ・ダーク	Drama	War	Drama	Romance

をうける映画がある．たとえば，IMDb によれば「タワーリング・インフェルノ」という映画のジャンルは「Drama」となっているが，この映画は，高層ビルでの火災をもとにしたパニック映画であり，ナイーブ・ベイズ分類による 1 位「Action」，3 位「Thriller」はこの映画のジャンルとしてもっともらしい．ほかにも，「仕立て屋の恋」はその名のとおり，仕立て屋が，ある女性に恋をしてしまい，悲劇にいたる映画である．この映画はナイーブ・ベイズ分類によると，「Drama」，「Romance」である．

これらの映画の，コメントに出現する単語に注目することで，ナイーブ・ベイズ分類がなぜ IMDb で付与されているジャンルと異なる答えを出しているのかがわかる．

例として，「タワーリング・インフェルノ」によせられているユーザのコメントの一部を(図 2)に示す．

(図 2) 「タワーリング・インフェルノ(1974／米／Drama)」のコメントの一部

★4 パニックものはどんなに役者を出したって，災害現場が主役なんだよ･･･とんでもない！ 豪華競演が面白いんですよ，この映画は．

★5 まったくもってその通りです．

★4 70年代にブームとなったオールスターによるパニック映画の中でも，群を抜いて面白い作品．ポール・ニューマンとスティーブ・マックイーンが同じ画面の中に収まってるだけで興奮してしまう

★5 結構今見ると安っぽい部分もあるし大味な作りなんだけど，それでも十分楽しめます．往年の大スター競演もいい感じです．初見時にTVに釘付けになった記憶がありますね．

★5 25年も前の作品だと思うとすごい．よくできてる．いい男2人も良い．「お父さんは心配症」でこの映画のネタあったなぁ･･･

★3 「コストを減らしたければ階数を減らせ」

★5 この映画のせいで，随分長いこと「フレッド・アステア＝上手い脇役」と思ってました．すんません．

★4 初めてみたパニック映画ってこれだったような･･･だいぶ記憶が飛んでるけど

★4 パニック映画として，ひさしぶりに考えさせられる良作

★4 あの状態で，イスで窓を破っちゃいけない…初めて知りました．ありえるからホラーより怖い．

★5 パニック映画の代表作 これと『ポセイドン・アドベンチャー』が双璧．でもね．．．

★5 子どもの頃見て，火災と高さに恐怖した．高層ビルははしご車が届かないとこの映画で納得．以後，はしご車の届く階にしか上らない･･つもりだったが，無理．

★4 夏休みの工作に「動くタワーリングインフェルノ」の巨大模型をつくって，デカすぎて持ってけなかったのは，私です．

コメントを単語ごとに切り出すと(表 7)のようになる．「タワーリング・インフェルノ」では，ナイーブ・ベイズ分類によると 1 位に「Action」，2 位に「Drama」，3 位に「Thriller」である．この映画に対するコメントをみて，直感的に，「パニック映画」「汗」「災害」「高層(ビル)」という単語がそれらのジャンルの順位に影響しているのではと仮定される．

(表 7) 「タワーリング・インフェルノ(1974／米／Drama)」コメント中に出現する単語(数字は出現回数)

映画	15	ニューマン	3	フレッド	2
見る	12	マックィーン	3	ポセイドン	2
作品	12	何	3	リメイク	2
する	11	火	3	印象的だ	2
パニック映画	9	汗	3	演技	2
ある	8	高層	3	於	2
頃	7	作る	3	価値	2
思う	7	子供	3	階	2
ない	6	初めて	3	感	2
の	6	大スター	3	記憶	2
ビル	6	知る	3	技術	2
マックイーン	6	怖い	3	詰込む	2
観る	6	おもしろい	2	競演	2
いい	5	これ	2	恐怖	2
いう	5	すごい	2	激突	2
できる	5	とき	2	見せ場	2
やる	5	はしご	2	減らす	2
人	5	ほど	2	言う	2
大作	5	ます る	2	娯楽	2
面白い	5	もう	2	後	2
いる	4	もる	2	豪華	2
なる	4	アステア	2	最近	2
もの	4	アドベンチャー	2	最高だ	2
パニック	4	インフェルノ	2	災害	2
今	4	オールスター	2	時代	2
良い	4	スター	2	車	2
こと	3	スティーブ	2	手	2
よい	3	テレビ	2	出来る	2
わかる	3	ドラマ	2	上	2

(表 8)「パニック映画」「汗」「災害」「高層」が各ジャンルに出現する回数

	パニック映画	汗	災害	高層
Drama	26	79	11	3
Action	37	69	21	3
Thriller	17	58	19	2
Comedy	1	35	0	0
SciFi	8	22	5	0
Adventure	13	20	5	0
Crime	0	19	2	0
War	0	16	0	0
Mystery	1	13	1	0
Romance	8	9	0	1
Animation	0	3	0	1
Horror	15	3	1	0
Documentary	0	2	0	0
Fantasy	0	2	0	0
Musical	0	2	0	0
Western	0	2	0	0
Family	0	0	0	0
Short	0	0	0	0

そこでこれらの単語が各ジャンルに含まれる映画にどの程度出現しているかを確認した(表 8). これにより, すべての単語について,「Action」,「Drama」,「Thriller」がほかのジャンルと比較して出現回数が多いことがわかった.

こういった単語が,「Action」的,「Thriller」的であるとされるならば,「タワーリング・インフェルノ」は「Drama」のみならず,「Action」,「Thriller」というジャンルに属する映画として, ユーザに提示されることは有意義なことだといえる.

次に「ときめきメモリアル」をとりあげる. これはアイドル女優たちが出演する, いわゆる学園ものの映画である. IMDb によると,「Drama」,「Romance」となっているが, ナイーブ・ベイズ分類は 1 位に「Horror」, 2 位に「Thriller」とまったく趣きの異なるジャンルに分類している.

実際のコメントを見ると, なぜこのようなジャンルに分類されるのかをうかがい知ることができる.「ときめきメモリアル」によせられたコメントを(図 3)にしめす.

(図 3)「ときめきメモリアル(1997／日／Drama・Romance)」のコメント

★4 このカメラワークと作品の爽やかさは『ダンサー・イン・ザ・ダーク』と対局を成す.
★3 ヤングジャンプ見てると思えばそれほど気にならない. ゲームと全然別物だった.
★3 これに3点(笑)！アイドル映画の醍醐味はあまりの寒気に背筋がゾクゾクっとする所. これはけっこう来ます.
★2 やっぱり女が見て楽しいものじゃなかった.
★2 アイドル好きの自分でも, 目を覆いたくなるようなシーンが続出. ある意味ホラーよりたちが悪い.

この映画は, ストーリーから考えれば, 正しいジャンルとして「Horror」や「Thriller」であるとはいえそうにもない. しかし(図 3)にも掲げた,『寒気に背筋がゾクゾク』,『ある意味ホラー』など, これらのコメントが「Horror」,「Thriller」という分類に影響を与えていることはあきらかである. したがって, 一般的には「Horror」,「Thriller」には属さない映画でも,「Horror」的,「Thriller」的との印象をうけたユーザがいるという情報を他のユーザに提供することができる. また, この例は, コメントを基にしたナイーブ・ベイズ分類が, 客観的・一般的な分類を行っているのではなく, よりユーザの主観性に近い分類を行っているということを示唆している.

6. まとめ

ナイーブ・ベイズ分類により, ユーザからのコメントに基づいた映画の分類が可能であることがわかった. さらに, IMDb による分類とナイーブ・ベイズ分類との間で結果が異なる映画に関しても, ユーザにとって有意義な情報を提示できる可能性を見出した.

しかし, ナイーブ・ベイズ分類による各ジャンルの順位が「あながち間違ってはいない」と言い切るにはその裏づけが必要である. そのために, 今後, 映画を分類するユーザアンケートを実施する. 多くのユーザに実際に映画を分類してもらい, それに基づき各映画でジャンルの順位付けをおこなう. このアンケートによる順位とナイーブ・ベイズ分類による順位を比較し, 評価する.

さらに, ナイーブ・ベイズ分類によるジャンルの順位情報をユーザに提示するシステムを開発する. IMDb による分類とナイーブ・ベイズ分類による上位ジャンルに大きな相違があれば, 影響している単語を抽出し, その単語を含むコメントをユーザに提示する. こうすることで, ユーザは映画についてより多角的な情報を得ることができると考える.

文　献

[1] Amazon.co.jp
http://www.amazon.co.jp

[2] アスクユー・レストランガイド
http://www.asku.com/rgj/

[3] PTP -Power to The People-
http://www.ptp.co.jp/

[4] CinemaScape
http://cinema.media.iis.u-tokyo.ac.jp/

[5] 木本晴夫, 特集:情報検索の新潮流 マルチメディア検索技術, 情報の科学と技術, Vol.50, No.1, pp.14-21, 2000.

[6] A. McCallum, K. Nigam, A comparison of event models for naive bayes text classification, Proc. of the AAAI-98 Workshop on Learning for Text Categolization, pp.41-48, 1998.

[7] The Internet Movie Database (IMDb)
http://www.imdb.com/

[8] allcinema ONLINE
http://www.stingray-jp.com/allcinema/

[9] ぴあシネマクラブ
http://www.pia.co.jp/cinemaclub/main.jsp

[10] 舘村純一, "協調型情報探索を支援する仮想評者とその視覚化", インタラクティブシステムとソフトウェアVII, 日本ソフトウェア科学会, pp. 147-152, 近代科学社, 東京, 1999.

[11] T. Joachims, Text categorization

[12] R. L. Rivest, Learning decision lists, Machine Learning, Vol.2, No.3, pp.229-246, 1987.

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

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