コメント文を利用する映画ナビゲーション

阿部 倫子（慶應義塾大学文学部図書館・情報学科, E-Mail:abebe9@slis.keio.ac.jp）
細野 公男（慶應義塾大学文学部図書館・情報学科, E-Mail: hosono@slis.keio.ac.jp ）
中川 裕志（東京大学情報基盤センター, E-Mail: nakagawa@r.dl.itc.u-tokyo.ac.jp）

1　はじめに

映画情報探索のための新たな方法を考案した．多数のユーザによって，映画に与えられている文章による評価（コメント）を映画間の関連性（類似度）の抽出に用いる，というものである．

2　コメントによる類似度とその意義

2.1 言葉による評価の類似性

映画推薦情報媒介システム CinemaScape [1)] では，映画に対してのユーザの感想や評価などを，コメントとして収集している．それらのコメントはデータベースに蓄積され，ユーザは各映画の情報と同時にそれらのコメントを参照することができる．

コメントは様々なユーザの主観的な映画に対する評価または感想，感情の表現といったもので成り立っている．映画間のこのコメントの類似度を算出すれば，映画どうしの「コメントが類似する」という意味での関連性を取り出し，映画の探索に利用することができる．つまり，言葉による評価が類似する映画をユーザに提示することができるということである．

2.2 既存システムとの違い

映画の探索に，ユーザの映画に対するコメントによる評価を用いることの意義は，「システムを作る側の恣意性が介入しない，人々のより自然な感性を映画間の関連性を構築することに用いることができる」というところにある．

テキストデータではなく，映画のようなマルチメディア型とよばれるデータには，従来の文献検索システムで考えられているような自動索引を行うことや，全文検索を行うことは不可能である．さらに，第三者の客観的な映画の分類を行うことも，エンターテイメント性の高いものには，あまり有意義であるとは考えらない．

また，ユーザの映画に対する採点による評価を映画間の関連づけに用いるという，協調フィルタリングを使ったシステムは，これらの問題を回避していることで注目されている．しかしこういったシステムの大きな問題点は，ユーザからの評価として用いられている，その採点という方法が，よいかわるいかという1次元的な尺度でしかない，ということである．

これに対し，コメントを分析しその類似度による関連性を利用すれば，ユーザの映画に対するより多様で詳細な評価を抽出することができる．

3 類似度計算の方法

コメントの類似度計算の流れは以下の通りである。

①コメントデータを形態素解析し、単語に切り出す。
②TF＊IDF により各単語に重み付けをする。
③ベクトル空間型モデルにより映画間の類似度を計算する。

まず映画ごとのコメントデータの全文を日本語形態素解析システム JUMAN3.61 を用いて処理した。今回は切り出した単語の中から、名詞、動詞、形容詞、形容動詞、副詞、未定義語 [2)]のみを利用することにした。次に、TF＊IDF によって映画ごとにこれら取り出した単語に重みをつける。さらに，これらの単語を軸にし、映画をベクトルに

するようなベクトル空間を作った。

このようにして映画間のコメントの類似度がそのベクトルのコサイン値として得られる(表1).

表1 コメントの類似する映画 (例:グーニーズ)

	グーニーズ	類似度
1	E.T.	0.338771
2	スタンド・バイ・ミー	0.282558
3	できごと	0.247013
4	ライフ・イズ・ビューティフル	0.230298
5	天空の城ラピュタ	0.2291
6	夢のバスにのって	0.228492
7	インディラ	0.227065
8	ニュー・シネマ・パラダイス	0.22685
9	バック・トゥ・ザ・フューチャー	0.222095
10	グレムリン	0.214268

4 結果・考察

コメントを用いて映画の類似度を得ることで,最も問題になるのは,コメントの数が映画によって異なっているということである.

1つもコメントを得ていない映画は,当然,全体の関連の中にとりいれることはできない.また,コメントの少ない映画が,映画間の類似度に対し無意味なノイズを生み出すとも考えられる.

そこで,1つの映画あたりに,どの程度のコメントを得ていればよいのか,という目安を探ることにした.

コメントが多い場合は望ましい類似度を得ている,という仮定のもとに,コメントを人為的に削減して,もっともコメントの多い状態で計算した映画間の類似度と,少しずつ削減した場合の類似度との相関がどう変化するかをしらべた.

今回の実験にはCinemaScapeにおけるコメントを用いたが,その概要を以下に示す.(数字は2000年11月現在.)

■ 映画の評価に用いられた単語
- 170234語/25313コメント/映画4937件

■ 1つのコメントから得られる平均的な単語数
- 7〜10語

■ コメントの数と映画数
- データベース中の全映 7939件
- 1コメント以上ある映画 4937件
- 5コメント以上ある映画 1357件
- 10コメント以上ある映画 647件
- 15コメント以上ある映画 396件
- 20コメント以上ある映画 255件

相関の変化を調べた結果,1つの映画につき7コメントまでコメントの数を減らしたところで,最もコメントの多い状態での類似度との正の相関関係は限界であるということがわかった(図1).

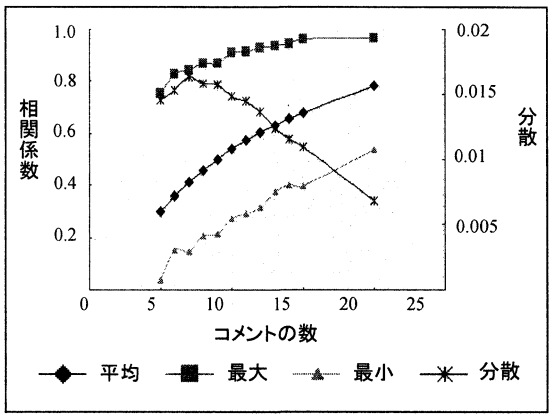

図1 コメントの数を減らした場合の相関の推移

よって,今回の実験データの場合,7コメント以上得ている961件の映画を対象に類似度を計算すれば,望ましい類似度を得られるであろう,と結論づけることができる.

4 注・文献

1) CinemaScape
http://cinema.media.iis.u-tokyo.ac.jp/

2) 未定義語:JUMANにおける,形態品詞分類辞書に登録されていない語。

3) 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WISS'99. 東京,近代科学社,1999, p.147-152,(レクチャーノート/ソフトウェア学23)