D-013

印象語の精錬による映画推薦システムの構築

Construction of the movie recommendation system by the refinement of the impression word

守屋 大地† 土屋 誠司‡ 渡部 広一‡
Moriya Daichi Seiji Tsuchiya Hirokazu Watabe

1. はじめに

近年，急速な情報社会の発展によって，社会には大量の情報が存在している．その中から，ユーザの力だけで必要な情報を探し出すのは非常に困難である．また，スマートフォンの普及や動画配信サービスの普及に伴い，気軽に映画を見られる環境が整ってきている．しかし，膨大な数の映画が存在しているため，個人の好みに合った映画を選ぶことは困難である．そのために個人の嗜好にあった映画を推薦する映画推薦システムを構築する．

映画を探す際，人はあらすじを見て視聴映画を決めることがある．本研究では，映画の要素として印象を使用することで、嗜好を抽出し，嗜好に合った映画を推薦する．また，最適な印象語を関連度計算に基づき抽出することで，印象語の精錬を行い，よりユーザが求める映画を推薦できるシステムを構築する．

2. 関連技術

本研究では，映画のあらすじから形態素解析により抽出した名詞・動詞と印象語知識ベースに格納された印象語との関連の強さを定量的に表す手法である関連度計算方式を使用する。

2.1 概念ベース

概念ベース[1]は，電子化された国語辞書などから機械的に構築された知識ベースである．概念ベースは，概念と定義された概念の意味特徴を表す単語（属性）から構成されている．

2.2 関連度計算方式

関連度計算方式[2]とは，概念ベース内の概念間の関連の強さを定量的に評価するものである．関連度の値は 0.0 から 1.0 の実数をとり，1.0 に近づくほど関連が強いことを意味する．

3. 映画推薦システム

まず，対象とした 135 作品の映画の中で視聴した映画の入力を行う．次に，元々格納されている 188 語の印象語と全映画との関連度を求めることで，印象語知識ベースの精錬を行う．その後，視聴した映画の入力を基に映画の嗜好を抽出する．さらに，抽出した嗜好をもとに未視聴映画に点数付けを行う．そして，点数が高い未視聴映画上位 5 作品を出力するシステムである．提案システムの流れを図 1 に示す．

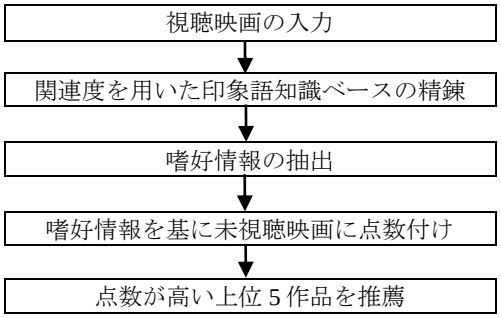

図 1 提案システムの流れ

3.1 印象語知識ベース

印象語知識ベースには，「映画の印象を表すのにふさわしい形容詞」が 188 語格納されている．格納されている形容詞は，TSUTAYA on-line[3]内の映画レビューで使われている形容詞のうち，大学生 12 名にアンケートを行った結果，被験者が見たことのある映画の印象を表現するために多く使用された形容詞である．表 1 に印象語知識ベースの例を示す．

表 1. 印象語知識ベースの例

印象語	愛らしい	温かい	・・・

3.2 印象語フィルタ

印象語フィルタは映画のあらすじ文から印象語を抽出するフィルタである．

まず，映画のあらすじ文を形態素解析し，名詞と動詞を抽出する．次に，抽出した名詞・動詞と関連度の高い印象語上位 10 語を印象語知識ベースから抽出する．

3.3 映画知識ベース

映画知識ベースとは，本研究で対象となる映画 135 作品のタイトルと印象という概念が格納されている知識ベースである．映画知識ベースを作成するために，映画のタイトルとあらすじを，TSUTAYA on-line[3]から収集する．収集したあらすじから印象を獲得し，映画知識ベースに付与する．表 2 は映画知識ベースの例である．

表 2 映画知識ベースの例

タイトル	印象
スパイダーマン	逞しい(0.0017) 優しい(0.001) ：

3.4 映画評価データ

映画評価データはインターネット上で 817 人に対してアンケートを行い取得したデータであり,135 本の映画それそ

† 同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡ 同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

れぞれに以下の 5 段階の評価が 817 人分記載されている．

①：過去（2 年より前）に観た
②：最近（2 年以内）観た
③：観たことはないが，今後観てみたい
④：観たことはなく，今後観てみたいか観てみたくないか分からない
⑤：観たことはないし，今後も観てみたいとは思わない

また，この映画評価データを用いて，5 段階評価の①②が付与された映画をあるユーザが視聴した映画，③④⑤が付与された映画をあるユーザの未視聴映画として扱う．

3.5 視聴映画の抽出

映画評価データからユーザの視聴履歴を獲得する．映画評価データ内でユーザが，①②の評価をした映画を，視聴済み映画をして獲得する．

3.6 印象語知識ベースの作成

本システムでは，135 作品全ての映画のあらすじ文から抽出した動詞・名詞と印象語知識ベースに格納された 188 語の印象語それぞれとの関連度を求めることで映画推薦に使用する印象語を降順にソートする．その降順にソートされた印象語を印象語知識ベースに付与することで印象語知識ベースの作成を行う．

3.7 嗜好情報の抽出

視聴済み映画にはユーザの嗜好情報が含まれていると仮定し，ユーザの映画評価データから取得した視聴映画を基に嗜好情報を抽出する．映画知識ベースから，あるユーザが視聴した映画に付与された印象語を獲得し，獲得回数の多い印象語を嗜好情報とする．また，印象の嗜好傾向を推薦に反映するため，嗜好情報として獲得した印象語に，獲得回数の割合を重みとして付与し，点数付けに用いる．

3.8 未視聴映画に対する点数付け

映画評価データ内でユーザが，③④⑤の評価をした未視聴映画に対して，嗜好情報を基に点数付けを行う．まず，ユーザの嗜好情報として抽出した印象語が付与されている未視聴映画を，映画知識ベースからすべて取得する．次に，嗜好情報の印象語の重みと，取得した映画に付与されている同じ印象語の重み同士を掛け，それらの値を足し合わせたものを映画の点数とする．嗜好情報として抽出した印象と一致する印象が 1 つしかない映画については，その印象についてのみ同様の点数計算を行う．

4. 評価

映画の印象を表すのにふさわしい印象語を選定するために，印象語知識ベースに格納する印象語の数を変えて推薦を行い，精度を比較する．それにより，印象を用いた映画推薦システムにおいて，最もふさわしい印象語の利用語数を調べる．また，元々格納されている188語の印象語のうち，映画の印象としてふさわしくないと考えられる印象語が多数存在したため，被験者3人にアンケートを取り，不要であると考えた印象語を削除した．その結果，印象語は39語にまで絞られた．それにより作成された印象語知識ベースと，全映画と関連度が高い印象語上位39語より作成された印象語知識ベースを用いて，両者の精度比較を行った．

4.1 評価手法

映画評価データを用いて推薦を行い，評価実験を行った．本システムでは，推薦された未視聴映画のうち，③を推薦されるべき映画，③または④の場合を推薦されてもよい映画とし，その割合で精度を評価する．

4.2 評価結果

精度評価を行うにあたり，印象語知識ベースに格納する印象語を全印象語 188 語から減らして実験を行った．具体的には，全印象語 188 語，全映画と関連度が高い印象語上位 150 語，上位 100 語，上位 50 語，上位 40 語を格納した印象語知識ベースを用いて実験を行った．印象語知識ベースに格納する印象語数ごとの精度の比較結果を表 3 に示す．188 語，150 語，100 語，50 語，40 語はそれぞれイメージ知識ベースに格納した印象語数である．

表 3. 評価結果

	188 語	150 語	100 語	50 語	40 語
③	26.19%	29.23%	29.72%	30.35%	30.58%
③または④	56.96%	59.14%	59.34%	59.83%	60.20%

また，アンケートにより絞られた印象語 39 語をシステム A とし，全映画と関連度が高い印象語上位 39 語をシステム B とする．システム A とシステム B の精度の比較結果を表 4 に示す．

表 4. 評価結果

	システム A	システム B
③	23.43%	30.35%
③または④	52.02%	59.85%

5. 考察

表 3 の結果から，映画知識ベースに格納する印象語数が少なくなるほど精度は上がり，全映画と関連度が低い印象語を削減していくことでシステムの精度が向上することが分かる．しかし，印象語知識ベースに格納する印象語数を 40 語より減らすと精度が下がった．印象語知識ベースに格納する印象語を減らしすぎることで，それぞれの映画の特徴を細かく表現できず，映画に区別をつけにくくなったことが原因であると考えられる．

また，表 4 の結果から，人間の感覚で不要であると判断し，削除することで残った印象語より，全映画のあらすじと関連度が高い印象語を使用した場合の方がシステムの精度は向上することが分かる．ある映画に対して誰もが持つ印象よりも，映画を見る人によって違う印象を持つため，人間の感覚で印象語を正確に削減することは難しいと考えられる．

謝辞

本研究の一部は，JSPS 科研費 16K00311 の助成を受けた．

参考文献

[1] 奥村紀之，土屋誠司，渡部広一，河岡司，“概念間の関連度計算のための大規模概念ベースの構築”，自然言語処理，Vol.14，No.5，pp.41-64，2007.

[2] 渡部広一，奥村紀之，河岡司，“概念の意味属性と共起情報を用いた関連度計算方式”，自然言語処理，Vol.13，No.1，pp.53-74，2006.

[3] TSUTAYA online，http://www.tsutaya.co.jp/index.zhtml