印象語フィルタによる映画推薦システムの提案

守屋大地[1]　渡部広一[2]　土屋誠司[2]

概要：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，膨大な数の映画から個人の嗜好にあった映画を選ぶのは困難であると考えられる．そこで，あらすじから印象を抽出する映画推薦システムを構築する．具体的には，映画の要素として印象を用いることで，視聴映画より抽出した印象が付与されている未視聴映画を映画知識ベースから取得する．その後，未視聴映画との関連性を算出し，作品を推薦する．

キーワード：推薦，嗜好

Proposal of movie recommendation system using impression word filter

DAICHI MORIYA[†1]　HIROKAZU WATABE[†2]
SEIJI TUCHIYA[†2]

Abstract: With the spread of the video distribution service, the environment where the movie can be easily viewed has been prepared. However, it is difficult to select a movie that suits individual tastes from a huge number of movies. Therefore, we build a movie recommendation system that extracts impressions from the synopsis. Specifically, by using impressions as movie elements, unviewed movies to which impressions extracted from viewed movies are given are acquired from the movie knowledge base. After that, the relevance to the unviewed movie is calculated and the work is recommended.

Keywords: Recommendation, RSeference

1. はじめに

近年，急速に情報社会が発展し，獲得できる情報量が膨大になっている．数ある情報の中から，ユーザの必要な情報を探し出すことは非常に困難である．映画についてこの問題を考える．公開される映画の年間本数は増加傾向にあり，過去に公開されたものを含めた映画の量は膨大である．また，Amazon primeやhuluなどの動画配信サービスの普及に伴い，気軽に映画を観る環境が整ってきている．動画配信サービスやレンタルビデオ店などを訪れると数多くの映画が存在し，その中から見たい映画を自力で見つけ出すことは困難であると考えられる．

この問題を解決するために行われているのが，推薦システムの開発である．推薦システムの方式には，協調フィルタリングと内容ベースフィルタリング[1]が存在する．協調フィルタリングとは，ユーザとその他のユーザの嗜好パターンの類似度に基づき，推薦を行う方式である．Amazonや楽天などの商用サイトで購買率向上のために用いられており，有効性が認められている．しかし，協調フィルタリングが機能するためには，多くの人がアイテムを評価する必要があり，利用者が少ないと使用できない問題点がある．一方，内容ベースフィルタリングとは，ユーザの行動履歴から嗜好パターンを推測し，検索対象の内容と比較して，ユーザが好むとシステムが判断したものを推薦する方式である．ユーザ以外の人の履歴を利用する協調フィルタリングより，ユーザ本人の履歴を利用する内容ベースフィルタリングの方が推薦システムの精度は高くなる．しかし，内容ベースフィルタリングはユーザの行動履歴に依存しており，ユーザの嗜好パターンを推測するには，推薦結果に対するフィードバックが必要となる．そのため，内容ベースフィルタリングには，ユーザにかかる負担が大きいという問題点があり，現在主流となっているのは協調フィルタリングである．そこで，内容ベースフィルタリングを用いつつ，ユーザへの負担が少ない推薦システムが必要であると考えた．そこで本稿では，内容ベースフィルタリングを基に，印象語フィルタという映画のあらすじ文から印象語を抽出するフィルタのアルゴリズムの見直しを行い，映画にふさわしい印象を用いた映画推薦システムを構築することで精度の向上を図る．

2. 概念ベースと関連度計算方式

概念ベースと関連度計算方式について述べる．

1 同志社大学大学院理工学研究科
Graduate School of Science and Engineering,Doshisha University.

2 同志社大学理工学部
Faculty of Science and Engineering,Doshisha University

2.1 概念ベース

概念ベース[2]とは，電子化された国語辞書や新聞記事などから機械的に構築された知識ベースである．様々な語（概念）が，それを特徴付ける語（属性）とその重要度を表す数値（重み）の対の集合によって定義されている．概念には属性とその重要性を表す重みが付与されており，約 9 万語の概念が収録されている．

ある「概念A」は，その概念と関連が深く，その概念の意味となると考えられる「属性 a_i」と，その属性の重要性を表す「w_i」の組の集合で表される．概念Aは以下の式(1)で表される．

$$概念A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_n, w_n)\} \quad (1)$$

概念ベースの特徴として，属性を成す単語群も概念ベースの中で概念として定義されている点がある．概念Aの意味定義を行う属性a_iを、概念Aの一次属性と呼ぶ．一次属性a_iを一つの概念と見なせば，a_iからさらにその一次属性を導くことができ，概念a_iから導かれた属性a_{ij}を元の概念Aの二次属性と呼ぶ．これを展開していくと，一つの概念Aは n 次属性まで持つことができる．概念ベースの具体的な例を表 1 に，概念ベースの構造を図 1 に示す．

表 1　概念ベースの例

Table1 Concept-based example

概念	属性
医者	(医師,0.34)(患者,0.11)(病院,0.08)・・・
病院	(医院,0.25)(手術,0.18)(施設,0.04)・・・
治す	(治療,0.43)(医療,0.21)(病気,0.13)・・・
⋮	⋮

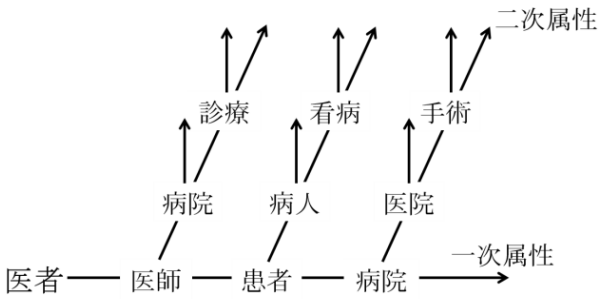

図 1　概念ベースの構造

Figure1 Concept-based structure

図 1 に示した通り，ある概念からは任意の次元までの属性を一次，二次，三次，…，N 次と導くことができる．このことより概念ベースは，N 次の属性による連鎖構造となっている．

2.2 関連度計算方式

関連度計算方式[3]とは，概念ベースに定義されている 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 の間の実数値で表され，概念間の関連が強いほど大きな値を示す．例えば，概念「自動車」に対して，「車」,「自転車」,「馬」の関連の強さを，表 2 に示すように数値化でき，コンピュータにも「自動車」と関連が最も強いのは「車」であると判断できるようになる．本研究では，関連度計算方式としてお互いの概念が持つ属性の一致度と重みを利用する重み比率付き関連度計算方式を使用する．

表 2　関連度計算方式の例

Table2 Example of relevance calculation method

基準概念	対象概念	関連度
自動車	車	0.92
	自転車	0.34
	馬	0.034

2.2.1 一致度

ある概念A，Bにおいて，その属性をa_i，b_j，各属性に対応する重みをu_i，v_iとし，それぞれ属性がL個，M個$(L \leq M)$とすると，概念A，Bはそれぞれ式(2)，(3)のようになる．

$$A = \{(a_1, u_1),\ (a_2, w_2),\ \cdots,\ (a_L, w_L)\} \quad (2)$$

$$B = \{(b_1, v_1),\ (b_2, v_2),\ \cdots,\ (b_M, b_M)\} \quad (3)$$

このとき，概念A，Bの属性一致度$DoM(A,B)$を式(4)，(5)のように定義する．

$$DoM(A,B) = \sum_{a_i = b_j} min(u_i, v_j) \quad (4)$$

$$min(u_i, v_j) = \begin{cases} u_i\ (u_i \leq v_j) \\ v_j\ (u_i > v_j) \end{cases} \quad (5)$$

ここで，$a_i = b_j$は属性同士が一致した場合を示している．つまり，一致度とは概念Aと概念Bそれぞれの属性の中で一致したものにおいて，小さい方の重みを選択して足し合わせた総和となる．これは，小さい方の重みは互いの属性の重みの共通部分であり，概念Aと概念Bどちらにも有効な重みであるためである．また，一致度を計算する際に各概念の重みの和が 1.0 となるように正規化する．したがって，一致度は 0.0 から 1.0 までの実数値をとる．

2.2.2 重み比率付き関連度計算方式

関連度計算方式では，奥村らの研究[4]により使用する属性は 30 個が適切であると報告されている．よって概念Aおよび概念Bの一次属性各 30 個を一致度が最大になるように組み合わせる．もし所持する属性数が 30 個に満たない概念の場合には，少ない方の属性数を計算に使用する個数としている．

2.2.1 項で述べた概念A, Bにおいて，まず属性数の少ない方の概念Aを基準とし，その属性の並びを固定する．その上で概念Bの属性を概念Aの各属性との一致度の和が最大になるように並び替える．並び替え後の概念Bの属性と重みを(b_{x_i}, v_{x_i})として式(6)のように定義する．

$$B = \{(b_{x1}, v_{x1}),\ (b_{x2}, v_{x2}),\ \cdots,\ (b_{x_M}, v_{xM})\} \quad (6)$$

これらの概念についての重み比率付き関連度$DoA(A,B)$は対応が決まった属性同士の一致度，重みの平均，重み比率の積として定義する．式は(7)のように定義する．

$$DoA(A,B) = \sum_i DoM(a_i, b_{xi}) \times \frac{(u_i + v_{xi})}{2} \times \frac{min(u_i, v_{xi})}{max(u_i, v_{xi})} \quad (7)$$

2.3 MeCab

MeCab[5]は日本語の形態素解析器の一つである．形態素解析とは日本語構造の制約を利用し単語の切り出しや品詞を同定することである．

形態素解析器として MeCab を使用する理由として，mecab-ipadic-NEologd が使用できることが挙げられる．mecab-ipadic-NEologd は多数の Web 上の言語資源から得た新語を追加した MeCab 用のシステム辞書である．MeCab の標準辞書である ipadic と mecab-ipadic-NEologd で「東京スカイツリーに行く」の形態素解析を行ったときの結果をそれぞれ図 2，図 3 に示す．

```
東京　トウキョウ　東京　名詞-固有名詞-地域-一般
スカイ　スカイ　スカイ　名詞-一般
ツリー　ツリー　ツリー　名詞-一般
に　ニ　に　助詞-格助詞-一般
行く　イク　行く　動詞-自立　五段・カ行促音便　基本形
EOS
```

図 2　MeCab の解析結果（ipadic）

Figure2 Analysis results of MeCab (ipadic)

```
東京スカイツリー　トウキョウスカイツリー　東京スカイツリー　名詞-固有名詞
に　ニ　に　助詞-格助詞
行く　イク　行く　動詞-自立　五段・カ行促音便　基本形
EOS
```

図 3　MeCab の解析結果（mecab-ipadic-NEologd）

Figure3 Analysis results of MeCab (mecab-ipadic-NEologd)

図 3 のように，標準辞書では「東京スカイツリー」という固有名詞を正しく解析することができないが，mecab-ipadic-NEologd を使用することで，図 3 のように正しく解析を行うことができる．本稿では，MeCab での形態素解析に使用する辞書として，2018 年 7 月 7 日時点の mecab-ipadic-NEologd を使用している．

3. 映画推薦システム

本章では映画推薦システムの定義や処理の流れ，使用技術について述べる．

映画推薦システムは，対象とした映画 135 作品の中から視聴映画に”1”を入力する．そして，視聴映画より嗜好を映画知識ベースから抽出する．最後に，抽出した嗜好を基に，未視聴映画に対して点数付けを行う．未視聴映画の中で点数が高い上位 5 作品を推薦する映画として出力する．推薦システムの流れを図 4 に示す．

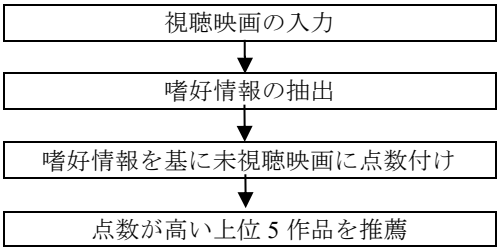

図 4　映画推薦システムの概要

Figure 4 summary of movie recommendation system

3.1 視聴映画の入力

視聴映画の入力は，135 本の映画のリストが格納された csv ファイルを用いて行う. ユーザは, 135 本の映画のうち，視聴したことのある映画それぞれに”1”を入力する. 視聴映画の入力例を図 5 に示す.

ガメラ 大怪獣空中決戦	
モンスターズ・インク	
トイ・ストーリー	1
パラサイト・イヴ	1
クレイマー、クレイマー	
水戸黄門 天下の副将軍	
風と共に去りぬ	
バットマン ビギンズ	

図 5　視聴映画の入力例

Figure 5 Example of watching movie input

3.2 映画知識ベース

映画知識ベースとは，本稿で対象となる映画 135 作品のタイトルと印象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 から収集する．収集したあらすじから印象語フィルタを用いることで印象語を獲得し，映画知識ベースに付与する．表 3 は映画知識ベースの例である．

表 3　映画知識ベースの例
Table3 Example of movie knowledge base

タイトル	印象
ガメラ 大怪獣空中決戦	怪しい(0.0258) 優しい(0.0096) ：
トイ・ストーリー	美しい(0.0097) 羨ましい(0.0073) ：

3.3 印象

映画知識ベースにおける印象は，印象語とその重みのセットで表される．印象語とは，映画の印象を表すのにふさわしい形容詞のことである．例えば，怪しいや優しい，美しいなどが挙げられる．それらの印象語は，名詞知識ベースを基に獲得される．

映画知識ベースに格納されている 135 作品には合計 23 ジャンルが存在する．印象語知識ベースには，TSUTAYA on-line[6]内からでジャンル別に 5 作品ずつ 135 作品以外の映画レビューを取得後,抽出した形容詞 176 語が格納されている．このようにして映画レビューから獲得した形容詞を印象語とする．

3.4 名詞知識ベース

名詞知識ベースは感情発生に関係する名詞と対応する 5 感感覚語・知覚語が登録されている．しかし，感情発生に関係する名詞は語数が膨大であるため，全てを知識ベースに登録することは困難である．そこで，名詞知識ベースには感情発生に関係する名詞のうち，代表となる 141 語（以下，象徴語）を対応する 5 感感覚語・知覚語と共に，名詞知識ベース内の象徴語知識ベースに登録している．入力された名詞が未知語の場合は未知語処理を行うことで，登録された象徴語と対応付けて 5 感感覚語･知覚語を取得する．未知語処理については 5.3 節で述べる．象徴語知識ベースの例を表 4 に示す．

表 4　象徴語知識ベースの例
Table4 Symbolic Knowledge Base Example

5 感感覚語・知覚語	象徴語
美しい	ルビー，絵画，珊瑚礁，美術品，花，紅葉，…
心強い	援軍，警備，防具，金庫，後ろ盾，同盟者，…

また，名詞によっては同じ分類でも感情発生の観点から分類すると異なる場合がある．例えば図 6 のように，「漫画」と「絵」は，一般的な観点では「絵」という分類でまとめることができる．しかし,「漫画」の 5 感感覚語・知覚語は「面白い」であるが,「絵」からは「面白い」という 5 感感覚語・知覚語は取得されない．この様に，同じ分類でも抱く感覚が異なり 1 つにまとめられないものがある．そこで，一般的には 1 つの分類にまとめられるが感情発生の観点からはその分類とは異なる名詞を代表語，代表語と一般的には同じ分類であるが感情発生の観点からは全く別の分類になる名詞を反例語とし，名詞知識ベース内の名詞代表語知識ベースに分類した 5 感感覚語・知覚語を登録している．名詞代表語知識ベースの例を表 5 に示す．

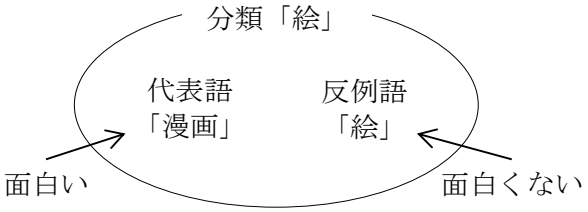

図 6　代表語と反例語の例
Figure6 Examples of representative words and counterexamples

表 5　名詞代表語知識ベースの例
Table5 Noun representative word knowledge base example

5 感感覚語・知覚語	代表語	反例語
美しい	孔雀 花火	鳥 火
面白い	漫画 海水浴	絵 海
うるさい	音痴 野次 蛙	歌 言葉 両生類

3.5 名詞の未知語処理

未知語処理とは，未知語と知識ベースに登録された代表的な語との意味的な近さを関連度として計算し，最大関連度の語に代替する処理である．

名詞が未知語の場合，未知語処理を行うことで名詞から 5 感感覚語・知覚語の取得を行う．処理の流れを図 13 に示す. 初めに, 名詞知識ベースの象徴語知識ベースを参照し，全象徴語と未知語との関連度を計算する．関連度の最大値が閾値（0.08）より高い場合は象徴語の 5 感覚語・知覚語を取得する．閾値以下の場合は名詞知識ベースの名詞代表語知識ベースを参照し，全代表語と未知語との関連度を計算して最大関連度を取得する．この最大関連度が閾値以上であれば，反例語と未知語との関連度を計算し，代表語と未知語との関連度と比較する．そして，代表語との関連度の方が高ければ代表語の 5 感感覚語・知覚語を取得する．例えば,「怪獣」という語は，未知語（概念ベースには登録

されているが，知識ベースには登録されていない語）である．名詞の未知語処理により，「怪獣」は「猛獣」に代替される．なお閾値には，複数の閾値の基に実験を行い，最も 5 感感覚語・知覚語の取得精度が高かった精度を採用している．

3.6 印象語フィルタ

TSUTAYA on-line から収集したあらすじと名詞知識ベースを利用して映画の印象を抽出する印象語フィルタを作成する. まず, 収集したあらすじの文章から名詞を抽出する．次に，抽出したすべての名詞を未知語処理により名詞知識ベースに含まれる名詞に代替する．そして，その名詞と対応付けられた知覚語を印象語とする．また，獲得回数の割合を印象語の重みとする．そして，獲得した印象語と重みをセットで映画知識ベースにおける印象とし，映画知識ベースに付与する．図 7 に印象語フィルタの流れを示す．

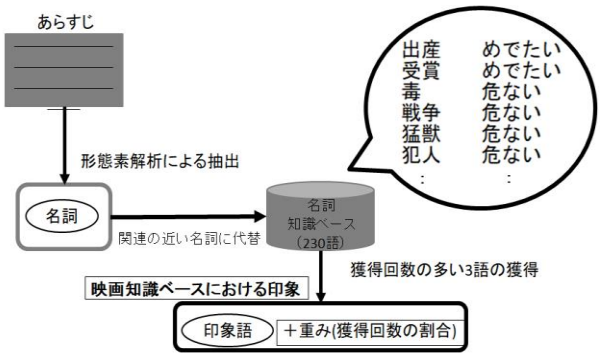

図 7　印象語フィルタの流れ

Figure7 Impression word filter flow

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. 抽出する嗜好情報の数については，6 章で説明する．

表 6 を用いて，嗜好情報抽出の例を示す．

表 6　嗜好情報の抽出

Table6 Extract preference information

視聴した映画	もののけ姫	ローマの休日	ショーシャンクの空に
印象語	美しい	美しい	美しい
	切ない	切ない	純粋な
	残酷な	幸せな	辛い

あるユーザの視聴した映画が「もののけ姫」「ローマの休日」「ショーシャンクの空に」の 3 本であったとする. まず, それぞれの映画に付与されている印象語を映画知識ベースから取得する. 取得した印象語が表 6 のようになったとし, 仮に抽出するユーザの嗜好情報を 2 つとすれば，多くの映画に付与された印象語である，「美しい」「切ない」が嗜好情報となる．また，「美しい」の重みは，3 本の映画すべてに付与されているので 3/3=1.00，「切ない」の重みは，3 本中 2 本の映画に付与されているので 2/3=0.66 となる．以上より, あるユーザの嗜好情報は,「美しい(1.00), 切ない(0.66)」となる．

3.8 未視聴映画への点数付け

未視聴映画とは，135本のうち視聴したことのない映画のことである．この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語と一致する印象語が1つしかない映画については，その印象語についてのみ同様の点数計算を行う．

表7を用いて，点数付けの例を示す．ここでは，映画知識ベースに付与する印象数は3つとして説明する．嗜好情報は3.6節の例で抽出した「切ない(1.00)，美しい(0.67)，怖い(0.67)」を使用する．

表7　点数付けのための映画知識ベースの例

Table7 Example movie knowledge base for scoring

タイトル	印象
トイ・ストーリー	切ない(0.5) 美しい(0.1) 意外な(0.04)

表7より，「トイ・ストーリー」には，先ほどの例で抽出した嗜好情報「切ない(1.00)，美しい(0.67)，怖い(0.67)」の，「切ない」「美しい」が付与されている．点数計算は，式(8)のようにして，それぞれの印象語について，嗜好情報としての重みと映画知識ベースに付与された重みとを掛け合わせて行う．「切ない」については，嗜好情報の重み1.00 と映画知識ベースにおける重み0.5を掛け，「美しい」については，嗜好情報の重み0.67 と映画知識ベースにおける重み0.1を掛ける．

$$1.00\times0.5 \quad + \quad 0.67\times0.1 \quad = \quad 0.567 \qquad (8)$$

この計算により，「トイ・ストーリー」の点数は 0.567 となる．

3.9 映画の推薦

嗜好情報を基に点数が付与された映画を降順にソートした後に，上位 5 作品を出力し．ユーザに映画を推薦する．図 8 システムの出力例を示す．

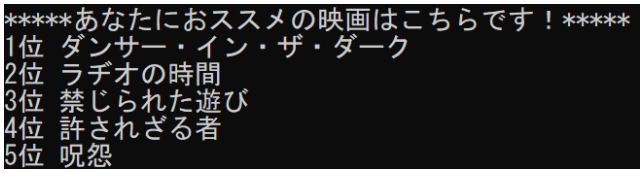

図 8 出力例

Figure8 Output example

4. 評価

4.1 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 作品の映画それぞれに以下の 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．

4.2 評価方法

映画評価データを用いて推薦を行い，評価実験を行う．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．

4.3 評価結果

映画推薦システムの精度を表 8 に示す．

表 8　評価結果

Table8 Evaluation results

	推薦されるべき映画	推薦されてもよい映画
システム	34.57%	63.62%

5. 考察

表 8 の結果から，映画推薦システムの精度は，推薦されるべき映画 34.57%，推薦されてもよい映画 63.62%となった．印象語フィルタを用いることで，映画から正確な印象語の抽出が可能になったことがこの精度に繋がったのではないかと考えられる. 例として,「ガメラ　大怪獣空中決戦」という映画の印象語の獲得例を表 9 に示す．

表 9　「ガメラ　大怪獣空中決戦」の印象語の獲得例

Table9
Acquisition example of impression words of
"Gamera Large Kaiju Aerial Battle"

映画名	獲得された印象語
ガメラ 大怪獣空中決戦	危ない
	怖い
	有害な

「ガメラ　大怪獣空中決戦」という映画は，タイトルの決戦という言葉からもわかるように，激しい戦いを繰り広げる映画である．表 9 から，印象語フィルタを用いることで，映画の印象を表すのにふさわしいと考えられる「危ない」や「怖い」といった印象語の抽出が可能になっていることが分かる．

6. おわりに

本稿では，印象語フィルタにより映画にふさわしい印象語の抽出が可能となり, 4.1 節の映画評価データにおける③の割合が最大 34.57%，③または④の割合が最大 63.62%を精度とする映画推薦システムを構築することができた．また，本システムでの入力は視聴履歴のみであり，内容ベースフィルタリングを用いた推薦システムの欠点である，ユーザの負担を軽減することができたと考える．

今後の課題として，さらに印象語を正確に映画に付与できる手法を考案することが鍵であると考える．この課題の解決策として，印象語フィルタのアルゴリズムを再検討することで，さらなる精度の向上が期待できる．精度が向上すれば，印象語を用いることで，現在主流となっている協調フィルタリングではなく，内容ベースフィルタリングを基にした映画推薦システムを実用化することも可能ではないかと考える．

謝辞　本研究を進めるにあたり，ご指導頂きました本学の渡部広一教授，土屋誠司教授に心から感謝致します．また，研究活動における諸問題の解決にご協力してくださった知識情報処理研究室の皆様に厚く御礼申し上げます．

最後に，あらゆる面で援助していただき，大学で研究を行うという貴重な場を与えてくださった両親，家族に心から感謝いたします．

参考文献

[1] Dietmar Jannach，Markus Zanker Alexander Felfering，Gerhard Friedrich ，“情報推薦システム入門”, 共立出版, pp.2-6, 2012.

[2] 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.

[3] 荒木孝允，奥村紀之，渡部広一，河岡司，“比較対象概念の共通属性を重視する動的関連度計算方式”，同志社大学理工学研究報告, Vol. 48, No. 3, pp. 14-24, 2007.

[4] 奥村紀之，荒木孝允，渡部広一，河岡司，“概念属性の動的評価に基づく概念関連度計算方式”，情報処理学会，E-033，pp.223-226，2006.

[5] “MeCab “，<http://taku910.github.io/mecab/>，(参照 2020-01-13).

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