ありきたりさを考慮した映画推薦アルゴリズムの評価

2015SC026 稲垣映奈　2015SC093 武井さやか
指導教員：河野浩之

1　はじめに

コールドスタート問題に対応可能なコンテンツベースで実装されるトピックモデルは映画ドメインなど特定のドメインでの使用はそれほど深く分析されていない [4]. また, コンテンツベースにおいて複数の特徴量を用いる場合すべてのアイテムがそれらの特徴量を持つとは限らない [2]. そこで映画コンテンツベースで映画のストーリーのみを特徴量とする推薦アルゴリズムを提案する. この提案アルゴリズムは, 入力された映画に対してストーリーの類似度は保ちつつもストーリーがありきたりではない映画を推薦する. ここで, ありきたりとはセレンディピティーとは違い新規性や有用性などユーザ側の評価は用いず, コンテンツ情報から類似する他のコンテンツの数に着目し相対的に判断されるものと定義する.

提案アルゴリズムの流れを以下に示す. はじめに映画サイトから抽出したストーリーの要約に対し tf-idf, LSA を行い映画間の類似度をコサイン類似度で求める. 次に, その類似度を用いてネットワークを形成する. 最後に, 入力された映画の作品名に対して, 高い類似度と低い次数中心性を持つ映画を推薦するように河合ら [3] の式の考え方を用いてスコア計算を行う. これにより入力された映画に対してストーリーの類似度は保ちつつも, ストーリーがありきたりでない映画を推薦する.

本稿は全 6 章で構成されており, 2 章では推薦システムの先行研究, 3 章ではありきたりさを考慮した映画推薦アルゴリズムの提案, 4 章ではありきたりさを考慮した映画推薦アルゴリズムの構築, 5 章ではありきたりさを考慮した映画推薦アルゴリズムの実験, 結果, 6 章ではまとめを示す.

2　推薦システムの先行研究

Sonia ら [4] はトピックモデルを用いて映画の推薦を行っている. 河合ら [3] はネットワーク伊藤ら [2] はアソシエーションルールを用いて意外性を考慮した推薦の研究をしている.

提案アルゴリズムで参考にした各先行研究をデータ, 手法, 特徴, の観点から比較した結果を表 1 に示す. 推薦アイテムの意外性という観点で見ると河合ら [3] はネットワークのハブの性質を用いることで対応し, 伊藤ら [2] はアソシエーションルールを用い独自の計算式を提案することで対応している. コールドスタート問題の観点で見ると伊藤ら [2] は対応できないが Sonia ら [4], 河合ら [3] はコールドスタート問題に対応する. また, LSA を使った推薦システムのサーベイによれば, LSA とネットワークのハブの性質を組み合わせた研究は紹介されていない [1].

3　ありきたりさを考慮した映画推薦アルゴリズムの提案

3.1 節で, ありきたりさを考慮した映画推薦アルゴリズムの概要, 3.2 節で, 映画間ネットワークの構造, 3.3 節で推薦スコアの計算方法について説明する.

3.1　ありきたりさを考慮した映画推薦アルゴリズムの概要

本節では, ありきたりさを考慮した映画推薦アルゴリズムの概要について説明する. 提案アルゴリズムでは, トピックモデルを用いた推薦システムの先行研究で扱われなかったストーリーのありきたりさに着目し, 類似度は保ちつつもありきたりでない映画を推薦する. 映画の類似度を求めるためには Sonia ら [4] の研究結果から, LSA とコサイン類似度を用いる. ありきたりでないストーリーの映画を求める手法には伊藤ら [2] と河合ら [3] の人気なアイテムほど推薦され難くすることで意外性を高めているという考え方を参考にし, 類似してる映画が多い映画ほど推薦され難くする. しかし, 伊藤ら [2] の研究では, コールドスタート問題に対処できない. よって, 河合ら [3] の研究に倣いネットワークとハブの性質を用いる.

また, 提案アルゴリズムにおいてありきたりの指標は, ストーリーの類似度によって形成されるネットワークにおいて次数中心性で表す.

図 1 のありきたりさを考慮した映画推薦アルゴリズムの概要図の (1) から (8) を以下で述べる. (1) 映画情報サイトから, スクレイピングとクローリングにより各映画のストーリーの要約を抽出する. 抽出された要約のテキストデータは, 形態素解析を行い, ストップワードを取り除いてから, 固有名詞以外の名詞, 動詞, 形容詞, 形容動詞の単語をキーワードとしてデータベースへ格納する. (2) 格納されたデータから, 各映画のストーリーの要約における各キーワードの重要度を L2 正規化した tf-idf によって計算し, それらを成分とするキーワード-文書行列を作成する. (3) キーワード-文書行列に対し LSA を行い, 文書-トピック行列を作成する. (4) 文書-トピック行列に対し, 文書間のコサイン類似度を計算することで映画間コサイン類似度を求める. (5) 求められた映画間コサイン類似度から, 映画間コサイン類似度が閾値以上のノード間でリンクを張り, 映画間ネットワークを形成する. (6) ユーザが入力した映画の作品名を受け取る. (7) 入力された映画に対して, 映画間ネットワークから求められる映画間の類似度と, 次数中心性を用いて推薦スコアを計算する. (8) 推薦スコアの高い順に映画を表示することで, ユーザが入力した映画とストーリーの類似度を保ちつつ, ありきたりでないストー

表 1 推薦システムの先行研究の比較

	データ	手法	特徴
[4]	ストーリー要約	トピックモデル (LSA)	コールドスタート問題に対応できる ストーリー類似度を用いて推薦を行う
[3]	購買履歴	商品ネットワーク	コールドスタート問題に対応できる ハブの性質を用いて意外性のある推薦を行う
[2]	ユーザ評価	アソシエーションルール	ユーザの間の嗜好の違いを示す計算式 を用いて意外性のある推薦を行う

リーの映画が推薦される.

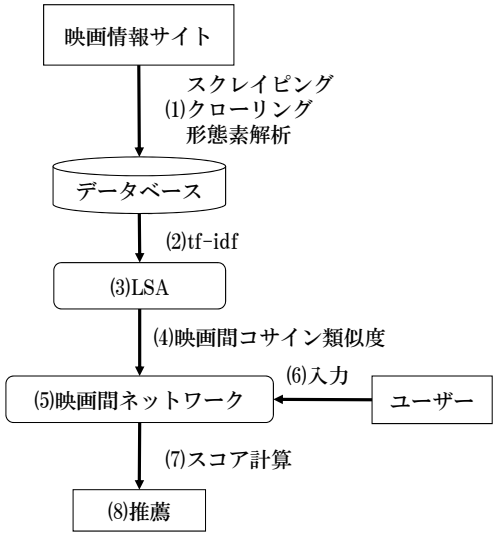

図 1 ありきたりさを考慮した映画推薦アルゴリズム概要図

3.2 映画間ネットワークの構造

本節では, 図 1 の (5) の映画間ネットワークについて詳しく説明する. 映画間ネットワークの構築に用いる以下の式 (1), (2), (4) は河合ら [3] のネットワークの式である. 河合ら [3] のネットワークでは, 商品をノードとし相関係数によってリンクを張ったが, 提案アルゴリズムでは, 映画をノードとし映画間コサイン類似度によってリンクを張る. また, ネットワーク上の類似度の定義は河合ら [3] の定義を用いる. ありきたりでない映画を推薦するためにはハブの性質を用い, その中でも河合ら [3] の定義したハブ度ではなく次数中心性を用いる.

映画間ネットワークは重み付き隣接行列 A で表し, 式 (1) のようになる.

$$A = [a_{ij}] \tag{1}$$

A の要素 a_{ij} は映画 i と j の間にリンクがあるときに正の値をとり, リンクがないときには 0 をとる. リンクは図 1 の (4) で求めた映画間コサイン類似度が閾値 th 以上の映画間に張る. ただし, th は実験により F 値が最大となる値を採用する. 映画間コサイン類似度行列 C は映画 i と j の映画間コサイン類似度 c_{ij} を要素とした行列であり, 式 (2) のようになる.

$$C = [c_{ij}]_{|D|\times|D|} \tag{2}$$

a_{ij} は式 (3) のように定義する. ただし, (i), (iii)-(v) は河合ら [3] によって定義されたものである. a_{ij} は映画 i と映画 j が等しいとき ((i) の条件), 映画 i, j の映画間コサイン類似度が負の値のとき ((ii) の条件), または映画 i, j の映画間コサイン類似度が閾値 th より小さく映画 i, j がそれぞれ映画 i, j 以外の映画とリンクを持っているとき ((iii), (iv) の条件) に 0 をとり, それ以外のとき ((v) の条件) は映画間コサイン類似度の値をとる. (iii), (iv) の条件は孤立ノードの発生を少なくするために付けた条件である.

$$a_{ij} = \begin{cases} & \text{if } i = j \quad \text{(i)} \\ 0 & \vee c_{ij} < 0 \quad \text{(ii)} \\ & \vee c_{ij} < th \wedge c_{kj} \geq th \ \exists k \neq j \quad \text{(iii)} \\ & \qquad \wedge c_{ik} \geq th \ \exists k \neq i \quad \text{(iv)} \\ c_{ij} & \text{otherwise} \quad \text{(v)} \end{cases} \tag{3}$$

ネットワークを辿ることで間接的にもストーリーの類似度が保たれている映画を推薦するため, 河合ら [3] に倣い映画間の最短経路における映画間コサイン類似度の積を映画間ネットワーク上の映画間の類似度とする. 以下, これをネットワーク類似度とする. 最短経路は Dijkstra 法を用いて求め, 類似度が大きいほどネットワーク上の距離を短くするために経路の長さは A の要素の逆数を用いて $\sum_{(v,w)\in P_{ij}} {a_{vw}}^{-1}$ とされている. 映画 i と j のネットワーク類似度 s_{ij} は式 (4) のように定義されている. P^*_{ij} は映画 i と j の最短経路であり, s_{ij} は 0 から 1 までの値を取る.

$$s_{ij} = \prod_{(v,w)\in P^*_{ij}} a_{vw} \tag{4}$$

提案アルゴリズムでは, リンクが少ない映画つまり次数中心性が低い映画は, ストーリーの類似している映画の数が少ないのでストーリーがありきたりでないと考えられる. 映画間ネットワーク上での映画 j の次数中心性 h_j は, 式 (5) のようになる. ここで, n は映画間ネットワークに含まれる映画の数である. b_{ij} はリンクの有無を表し, a_{ij} が 0 より大きいとき 1 を取り, それ以外は 0 である. h_j は 0 以上の値を取る.

$$h_j = \sum_{i=1}^{n} b_{ij} \tag{5}$$

3.3 推薦スコア計算

本節では, 図 1 の (7) の推薦スコア計算について詳しく説明する. 提案アルゴリズムでは, 入力された映画のストーリーとの類似度を保ちつつありきたりでない映画を推薦するため, ネットワーク類似度が高く次数中心性の低い映画を推薦する.

よって河合ら [3] の推薦スコアの考え方に倣い, 映画 i に対する映画 j の推薦スコア r_{ij} をネットワーク類似度 s_{ij} とパラメータ λ, 次数中心性 h_j から式 (6) と定義する. ここで, λ は h_j の重みを調節するパラメータであり, λ が小さいほどよりありきたりでない映画を推薦すると予想される.

$$r_{ij} = s_{ij} + \lambda h_j \tag{6}$$

4 ありきたりさを考慮した映画推薦アルゴリズムの構築

本章では, ありきたりさを考慮した映画推薦アルゴリズムの構築について説明する. まず実装環境について説明し, 次に映画間ネットワークと推薦スコア計算式の構築について説明する. 実装には, スクレイピングツール, クローラとして Ruby の Nokogiri, Anemone を使用し, LSA の計算, ネットワークの構築, パラメータの最適化には Python の gensim, NetworkX, scikit-learn を用いる. また映画情報サイトには, あらすじやレビュー, ランキングなどの情報が記載されている, Yahoo!映画, Filmarks, 映画.com, また, ストーリーのネタバレ記事が記載されている映画ウォッチ, hmhm などがある. 提案アルゴリズムでは, Sonia ら [4] と同様に映画全体の内容を反映させたデータをストーリーの要約として用いるため, ネタバレ記事を使用する. そこで, ネタバレ記事があり, よりデータ数が豊富な映画ウォッチ (https://eiga-watch.com/) を選択する.

映画間ネットワークの構築には, まず, NetworkX からネットワークを生成するモデルを読み込む. 次に, LSA とコサイン類似度によって求まる各映画間の映画間コサイン類似度から, 閾値以上の映画間コサイン類似度を持つ映画間をモデルに格納しリンクを持たせる. ただし, 閾値以上の映画間コサイン類似度を持つ相手がいない映画は, そのうち 0 より大きく最大の映画間コサイン類似度を持つ相手とリンクを張る.

推薦スコア計算式の構築には, ある映画から他の全ての映画に対する最短経路における映画間コサイン類似度の総乗 (ネットワーク類似度) を計算する関数を定義し, その関数と NetworkX の次数を求めるコマンドを映画間ネットワークに使用し 3.3 節の式 (6) を作成し計算する.

5 ありきたりさを考慮した映画推薦アルゴリズムの実験

本章では, ありきたりさを考慮した映画推薦アルゴリズムの実験について説明する. まず, パラメータの最適化実験と提案アルゴリズムの性能評価実験を行い, さらに推薦結果の考察を行った. パラメータ最適化実験は scikit-learn により F 値が最大となる値を求め, 実験結果から LSA の次元数は 40, 映画間ネットワークのリンクの閾値は 0.4 とする. ここでは, 実験の詳細については割愛する.

性能評価実験では, それらの値を用いて実装した提案アルゴリズムにおいてパラメータ λ を-0.1 から 0 の範囲 (-0.1 $\leqq \lambda \leqq$ 0) で変化した時の推薦結果における類似度を保っている割合と次数中心性の平均値を計算し, その結果を丸・実線 (類似度を保っている割合) とバツ・破線 (次数中心性の平均値) のグラフで図 2 として示す. ここで, 推薦結果の類似度を保っている割合とは推薦結果の 10 個の映画のうち入力元の映画との映画間コサイン類似度が 0 より大きい映画の割合であり, 次数中心性の平均値は推薦結果の映画の次数中心性を平均した値である. 実験は映画ウォッチから抽出した 6861 件のネタバレ記事を使用した.

図 2 から $-\lambda$ がおよそ 10^{-3} よりも小さい時は類似度を保っていることがわかった. また, $-\lambda$ が大きくなるにつれ次数中心性は徐々に減少して行くことから, 当初の目標である類似度を保ちつつありきたりでない映画を推薦するには $-\lambda$ を 0 から 10^{-3} 付近の範囲に設定することで達成され, 10^{-3} に近づくほどありきたりでない推薦結果になると思われる.

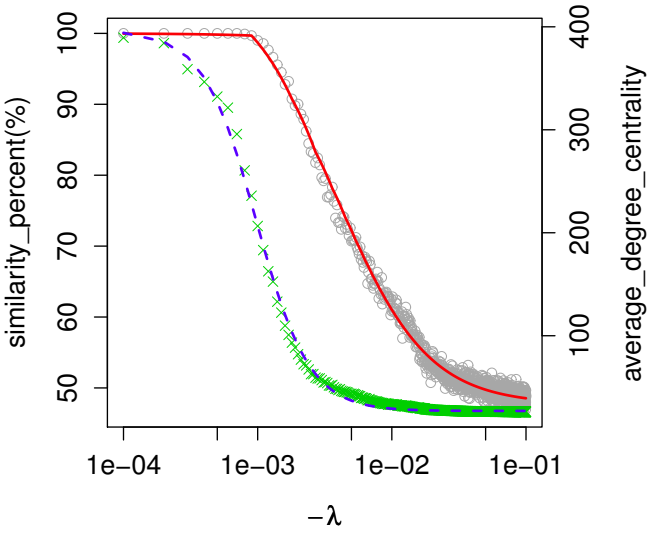

図 2 推薦結果の類似度を保っている割合と推薦結果の次数中心性の平均値

$-\lambda$ を 5×10^{-4} から 10^{-3} まで変化させ「白雪姫」を入力した時の推薦結果の変化を表 2 に示す. ここで [ ] の中の数値はその映画の次数中心性の値を示す. 表 2 で, $-\lambda$ が 10^{-3} のとき「オーロラ」などが新しく推薦されている. これは $-\lambda$ が 5×10^{-4} のときに推薦されていた次数中心性 428 の「美女と野獣 (2017 実写)」などに対して, 「オーロラ」の次数中心性が 342 なので, $-\lambda$ を大きい値に変更したことにより次数中心性が低い映画が上位に推薦されている.

また, 表 2 から $-\lambda$ が大きくなるに連れて次数中心性の

表 2 $-\lambda$ が 5×10^{-4}, 10^{-3} のときの「白雪姫」からの推薦結果

$-\lambda$	5×10^{-4}	10^{-3}
1	眠れる森の美女 [387]	五日物語 3 つの王国と 3 人の女 [360]
2	白雪姫と鏡の女王 [383]	眠れる森の美女 [387]
3	五日物語 3 つの王国と 3 人の女 [360]	シンデレラ (2015 年実写版) [324]
4	シンデレラ (2015 年実写版) [324]	白雪姫と鏡の女王 [383]
5	ナルニア国物語 第 2 章 カスピアン王子の角笛 [402]	プリンス・オブ・ペルシャ/時間の砂 [340]
6	シンドバッド 虎の目大冒険 [402]	アラジン [342]
7	プリンス・オブ・ペルシャ/時間の砂 [340]	オーロラ [342]
8	美女と野獣 (2017 実写) [428]	シュレック 3 [347]
9	塔の上のラプンツェル [414]	星の王子さま [337]
10	アラジン [342]	エバー・アフター [325]
次数中心性の平均値	378.2	348.7

平均値も小さくなり, $-\lambda$ が大きい値になるほどよりありきたりでない映画を推薦している. 例えば, 表 2 で $-\lambda$ が 10^{-3} のときに最初に推薦される「五日物語 3 つの王国と 3 人の女」は, お姫様やお城など「白雪姫」と酷似した設定を持ちながらも女性に眠る本性や欲望をテーマとしたダークファンタジーな内容となっている.

また, 実験結果から二つの課題が見つかった. 第一に, ある映画 A に対するネタバレ記事内の一部の固有名詞が MeCab により名詞と判断され削除されず, さらにその単語がネタバレ記事内で多用されていた場合 LSA による映画 A に対するトピック分布が映画 A のストーリーに則さないものとなる. よって映画 A に対するありきたりさや, 類似度の判断が想定されるものとは異なる場合がある. 例えば, 007 シリーズの映画の場合は主人公の名前であるボンドが人名と判断されず上位のトピックにボンドが含まれてしまったのでありきたりでないストーリーの映画だと判断されてしまっている. この課題に対して, 形態素解析の処理で映画特有の単語を固有名詞と判断させるために映画専用の辞書を作成することが解決策になると予想される.

第二に, ストーリーの設定のありきたりさを判断することは可能だがストーリーの展開に関してはコサイン類似度, LSA, ネットワークだけでは完全には網羅できない. 例えば「8 年越しの花嫁」を推薦元の映画とした場合, 結婚や病気などストーリーの設定は類似しているが恋人になってから病気になる展開と病気になってから恋人になる展開の違いは判断できていない. これは, 提案アルゴリズムが単語の出現する順番を加味していないことが原因と想定される.

6 まとめ

本論では, 映画コンテンツベースでストーリーのみを特徴量とし LSA, ネットワーク, 次数中心性を用いてありきたりでない映画を推薦するアルゴリズムを提案した. また, パラメータ λ の変化による類似度, 次数中心性, 推薦結果の変動を見ることで, 実際に λ が有効なことを示し, さらに次数中心性が調節可能なことを示した. 実験結果から提案アルゴリズムにおいては類似度を保ちつつありきたりでない映画を推薦するという当初の目標を達成した.

提案アルゴリズムにおいての次数中心性, 類似度の一般性, 満足度を評価するためにアンケートを行う必要がある. 具体的には $-\lambda$ が 0, 5×10^{-4}, 10^{-3} の推薦結果上位 3 つの各映画に対してありきたりさ, 類似度を 5 段階で評価してもらう. その結果から次数中心性, 類似度の一般性を証明し, さらにユーザの 5 段階評価からありきたりさが低く類似度の高い映画ほどアルゴリズムに対するユーザの満足度が高いと判断する.

参考文献

[1] Hanafi, Nanna Suryana, Abdul Samad Bin, Hasan Bashari, “Paper Survey and Example of Collaborative Filtering Implementation in Recommender System,” Journal of Theoretical and Applied Information Technology, Vol.95, No.16, pp.4001-4014, Aug. 2017.

[2] 伊藤寛明, 吉川大弘, 古橋武, “アソシエーションルールを用いた推薦システムにおける精度と意外性の向上,” 情報処理学会論文誌, Vol.8, No.2, pp.10-22, July 2015.

[3] 河合未夢, 能上慎也, 保坂忠明, 安藤晋, “商品間のネットワーク構造を用いた意外性を高める推薦モデルの提案,” 情報処理学会研究報告, Vol.2016-MPS-107, No.15, pp.1-7, May 2016.

[4] Sonia Bergamaschi, Laura Po and Serena Sorrentino, “Comparing Topic Models for a Movie Recommendation System,” WEBIST 2014-International Conference on Web Information Systems and Technologies, No.2, pp.172-183, Apr. 2014.