D-019

概念ベースと ALS を用いた日付連動型映画推薦システム
A Date-Based Movie Recommendation System Using Concept-Based Similarity and ALS

横田 篤洸† 土屋 誠司‡ 渡部 広一‡
Atsuhiro Yokota Seiji Tsuchiya Hirokazu Watabe

1. はじめに

現在，多くの定額制動画配信サービスが存在し，ジャンルや年代を問わず膨大な数の映画が提供されているが，その反面，ユーザーが自身の嗜好に合った作品を選択することは容易ではなくなっている．

日付は，祝日や記念日，誕生日など，ユーザーの心理的・行動的変化と深く関係しており，映画選択のトリガーとなり得る要素である．日付に関連する出来事と映画作品との意味的な関連性を分析し，それに基づいた推薦を行うことで，ユーザーの視聴意欲の向上を図る．

本研究では語彙の違いや言い換えに対応可能な概念ベース[1]と EMD[1]による文間関連度計算を導入する．また，ALS（交互最小二乗法）による行列因子分解を導入することで，ユーザーの潜在的な嗜好をより柔軟かつ高精度に捉えることを目指す．

2. 関連技術

2.1 概念ベース

概念ベース[1]とは複数の国語辞書や新聞などから機械的に構築した単語(概念)とその意味特徴を表す単語(属性)の集合からなる知識ベースである．概念には属性とその重要性を表す重みが付与されている．

2.2 関連度計算

関連度計算[1]は概念の二次属性間の一致度計算により求めた値をもとに概念間の関連性を数値として算出する．任意の概念A，Bについて，それぞれ一次属性をa_i，b_jとし，対応する重みをu_i，v_jとする．このとき，概念A，Bの一致度$MatchWR(A,B)$を以下の式(1)で定義する．

$$MatchWR(A,B) = \sum_{a_i=b_j} \min(u_i, v_j) \qquad (1)$$

属性が一致した場合に，重みの大きい方は重みの小さい方の重みを引き，もう一度，ほかの属性と対応をとることにする．対応を求めて，対応の取れた属性の組み合わせ数をT個とした場合，概念A，Bの関連度$DoA(A,B)$を以下の式(2)により定義する．

$$DoA(A,B) = \sum_{i=1}^{T}\{MatchWR(a_i, b_{ij}) \times (u_i + v_{xi}) \times (\min(u_i, v_{xj})/\max(u_i, v_{xj}))/2\} \qquad (2)$$

2.3 EMD (Earth Mover's Distance)

EMD[1]は線形計画問題の一つであるヒッチコック型輸送問題において計算される距離尺度であり，2 つの離散分布において，一方の分布を他方の分布に変換するための最小コストとして定義される．

EMD を文書検索に適用するには需要地と供給地，需要量と供給量，各需要地と供給地間の距離を定義する必要がある． EMD は文書間が類似していると値が低くなり，類似していないと高くなる．よって値が低い文章から順にユーザーに提示することで文書検索が実現できる．

2.4 ALS (Alternating Least Squares)

ALS[2]は行列因子分解の一種であり，特に大規模な疎な評価行列に対して効率的に適用できるアルゴリズムである．

協調フィルタリングにおける行列因子分解は以下のようになっている．ユーザー×アイテムの評価行列をユーザーの潜在特徴行列とアイテムの潜在特徴行列に分解し，以下の式(3)ように内積で近似する．

$$\hat{r}_{ui} = \boldsymbol{p}_u \cdot \boldsymbol{q}_i^T \qquad (3)$$

ここで，$\boldsymbol{p}_u$はユーザー$\boldsymbol{u}$の潜在ベクトル，$\boldsymbol{q}_i$はアイテム$\boldsymbol{i}$の潜在ベクトルを表す．目的は観測された評価値r_{ij}に対して予測値$\hat{r}_{ui}$を近づけることである．

ALS のアルゴリズムは以下の式(4)の損失関数を最小化する．

$$L = \sum_{(u,i)\in R}(r_{ui} - \boldsymbol{p}_u \cdot \boldsymbol{q}_i^T)^2 + \lambda(\|\boldsymbol{p}_u\|^2 + \|\boldsymbol{q}_i\|^2) \qquad (4)$$

ALS はユーザー行列Pとアイテム行列Qを交互に固定して最小二乗法で反復更新を行う．

3. 提案手法

3.1 使用データ

3.1.1 日付データ

日付データは，日本語版の Wikipedia の API を通じて収集した 365 日分の日付に関するデータである．

各日付に対して出来事，誕生日，記念日の三つのカテゴリに分類される記述が含まれている．本研究では，この日付データを映画の概要文との関連度を計算するクエリ文として利用する．

3.1.2 映画データ

映画データは，日本で公開された映画作品のメタデータを用いる．具体的にはレビューサイトである TMDb[4]の API を用いて取得した日本での公開日が 1980 年から 2025 年の映画カテゴリに属する作品のタイトル，概要文，出演者，監督，脚本のデータである．映画の概要文を日付データとの文間関連度計算に使用する．

3.1.3 評価データ

評価データは, 2023 年 10 月に収集された約 3 億件のユーザーによる評価を含む Movie Lens 32M[5]というデータの中から映画データとマージ可能な映画 ID のみを抽出した

† 同志社大学大学院理工学研究科
‡ 同志社大学理工学部インテリジェント情報工学科

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

4517212 件のユーザーによる評価である.評価データは ALS による学習モデルにユーザーとアイテム(映画)の評価行列として用いる.

3.2 日付と映画の関連付け

日付に関連する記述と映画との関連度を計算するために，本研究では EMD を利用した手法を採用する．これはもともと文書間の関連度を評価する手法であり，需要地と供給地，需要量と供給量，各地点間の距離を定義することで応用可能である．

本研究では以下のように要素を定義する．需要地は日付に関連する Wikipedia 記述(出来事，記念日)から抽出された索引語，供給地は映画の概要文から抽出された索引語，需要量と供給量はそれぞれの索引語に対する tf・idf 重み，距離は概念ベースを用いて計算した索引語間の意味的な関連度と定義する．EMD が小さいほど，その映画はその日付の出来事や記念日と意味的に関連していると解釈できる．日付に関連する Wikipedia 記述それぞれに対し EMD が低い 5 本の映画のデータを映画データから抽出したデータを日付映画データと定義する．

3.3 ALS による学習

ALS はユーザーの潜在特徴ベクトルと映画の潜在特徴ベクトルのどちらか片方を固定し，もう片方の行列に対して最小二乗法で逐次最適化を行う．損失関数が収束または最大繰り返しの回数まで繰り返す．評価データを評価行列として用いて，予測評価値の高い映画の抽出を目的とする．

3.4 二段階フィルタによる推薦

概念ベースと EMD を用いた文間関連度計算により抽出された日付に関係する映画の中から，ALS による学習を通じて得られた予測評価値の高い映画を推薦する．この二段階のフィルタを用いることで日付と意味的に関連し，かつ嗜好に合致する映画を推薦することができる．

具体的なユーザーによる入力とシステムによる出力を述べる．まず，ユーザーが入力するのはユーザーが好きな映画のタイトルと好きな日付である．入力するタイトルの数は，10 以上とする．入力された好きな映画のタイトルをもとに ALS を用いた学習モデルでユーザーの潜在特徴ベクトルと映画の特徴ベクトルを決定し，未視聴映画の予測評価値を求める．また，入力された好きな日付に関連する映画を 3.2 項で定義した日付映画データセットから抽出する．入力された好きな日付に関連する映画の中で予測評価値が高い順に 5 本の映画がシステムの出力となる．

4. ALS を用いた学習モデルの評価手法

ALS によって学習された推薦モデルの有効性を検証するため，定量的な指標として RMSE(平均平方根誤差)および MAE(平均絶対誤差)を用いて評価を行った．RMSE の特徴は誤差の二乗をとるため，大きな誤差に対してより敏感である．MAE の特徴は，各誤差を等しく扱うため，予測の全体の一貫した精度を評価するのに適している． 潜在特徴数と正則化項はそれぞれ潜在特徴数を 10,20,30 の 3 通りと正則化項を 0.01，0.1，1.0 の 3 通りで合計 9 通りの組み合わせで学習し，もっとも RMSE と MAE の値が低くなったモデルを使用した．

評価対象として，ALS に加え，行列分解に基づく代表的な協調フィルタリング手法である SVD（特異値分解）および NMF（非負値行列因子分解）と比較を行った．SVD はユーザーとアイテムの評価行列を潜在特徴空間に射影することで推薦を行う手法であり，広く利用されている．NMF はすべての因子が非負であるという制約を課すことで，より解釈性の高い潜在特徴を学習できる点が特徴である．

5. ALS を用いた学習モデルの評価

代表的な行列分解ベースの協調フィルタリングモデルである SVD と NMF に対し，RMSE および MAE を指標として予測精度を測定した．結果を表 1 に示す．SVD モデルは最も低い誤差を示し，ALS モデルは僅差でそれに続いた．一方，NMF モデルは他の 2 手法に比べて誤差が大きい結果となった．

モデル	RMSE	MAE
SVD	0.8561	0.6550
ALS	0.8579	0.6656
NMF	0.9068	0.6960

表 1　学習モデル別の RMSE と MAE の比較

6. 考察

本研究で使用した ALS モデルは，潜在特徴数や正則化項を最適化することで，SVD と同等レベルの RMSE および MAE を達成した．これは，ALS が評価値のばらつきを吸収しやすく，大規模な疎なデータに対して過学習を抑えつつ安定した学習が可能である点が影響していると考えられる．さらに，ALS は非負値制約を適用できるため，ユーザーやアイテムの潜在ベクトルが現実的な意味を持ちやすく，そのことが予測値の安定性につながったと推察される．

7. おわりに

本研究では，日付に関係する記述と映画の概要文の文間関連度に基づく推薦システムの実現に向けて，ALS を用いた学習モデルの有効性を評価した．SVD や NMF との比較実験の結果，ALS は高い精度を維持しつつ，大規模データへの適応性にも優れていることが示された．今後は、未実装である日付情報と映画の関連度計算部分を統合し，記念日や出来事に対して自然な形で映画を推薦できるシステムの完成と，そのシステムに対するランキング精度指標を用いた定量的な評価と定性的評価を行う予定である．

参考文献

[1] 藤江悠五，渡部広一,河岡司 “概念ベースと Earth Mover’s Distance を用いた文書検索”, 自然言語処理，Vol. 16, No. 3, pp. 3_25–3_49（2009）.
[2] Y. Hu, Y. Koren, and C. Volinsky. “Collaborative Filtering for Implicit Feedback Datasets,” Proc. IEEE ICDM, pp. 263–272 (2008).
[3] TMDB, https://www.themoviedb.org/?language=ja, 2025/4/20
[4] Movie Lens 32M, https://grouplens.org/datasets/movielens/32m/, 2025/4/21

Copyright © 2025 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.