D-003

メタデータと映像特徴に基づく内容ベース映像推薦

Content-based Video Recommendation Based on Metadata and Video Features

吉田 大我† 入江 豪† 佐藤 隆† 東野 豪†
Taiga Yoshida Go Irie Takashi Satou Suguru Higashino

1．はじめに

視聴映像数の爆発的な増加に伴い，嗜好に合った映像を見つけることはますます困難になりつつある．この問題への対策の 1 つとして，ユーザの嗜好に合う映像を自動的に推薦する映像推薦技術が注目を集めている．履歴ベースの手法は，利用可能な履歴の量が多い場合には効果的な推薦を行えるが，そうでない場合には推薦精度が低下するという問題がある[1]．これに対し，我々は映像に付与されたメタデータ（タグ）のランキングに基づく内容ベースの手法を提案してきた[2]．この手法は履歴を利用しないため，履歴の量に関わらず効果的な推薦が可能であるという利点がある．しかしながら，タグを利用して推薦するという特性上，付与されているタグが少ない場合には推薦精度が低下するという問題があった．

この問題を解決すべく，本研究では，メタデータに加えて，映像自体を解析して得られた特徴量を用いた推薦手法を提案する．動画共有サイトの映像を用いた実験により，メタデータのみに基づく推薦に比べて提案手法の推薦精度が良いことを確認した．

2．関連研究

これまでにも同様の問題に着目し，メタデータ以外の情報を利用して推薦精度を改善する試みがなされてきている．外部から得られる情報を利用して精度を改善する試みとしては，Wikipedia(注 1)の情報に基づいて映像に付与されたタグを選定し，Wikipedia のカテゴリに紐付けることにより，推薦精度を向上させる手法がある[3]．しかしながら，依然として推薦に利用するメタデータが少ない場合には，精度が低下してしまうという問題がある．

本研究に類似する，映像特徴を利用した試みとしては，Yang[4]らの手法がある．この手法では，テキストに基づく類似度に加え，映像全体における色ヒストグラム，動きの大きさ，カットの頻度，音のテンポに基づく類似度を組み合わせて映像間の類似度を算出する．実験により，映像特徴に基づく類似度を組み合わせることによって，推薦精度が向上できることが示されている．この手法は，映像全体の映像特徴の平均や標準偏差を算出し，特徴量としており，各映像特徴の時間的変化を考慮していない．これに対し本研究では，さらに代表シーンの色情報および映像中のイベントの時間的な変化も分析することにより，推薦精度を向上する手法を提案する．

3．提案手法

本研究では，メタデータと映像特徴を組み合わせた推薦手法を提案する．メタデータに基づく推薦手法には，タグランキングの比較による推薦手法[2]を用いたため，説明は割愛する．以下では，映像特徴に基づく推薦手法の詳細について述べる．

映像推薦では，ユーザの嗜好に寄与する映像特徴を抽出することが重要である．本研究では，映像の中で印象に残りやすいと考えられる代表シーンの見た目，および，映像の雰囲気に関係すると考えられるカットなどのイベントの構造的な特徴という 2 つの特徴量によって映像を表現し類似度を算出する．

3.1 代表シーンに基づく類似度算出

映像には様々なシーンが含まれている．このとき，ユーザの嗜好に影響を与える度合いはシーンによって異なると考えられる．本研究では，長い時間表示されるシーンが映像の代表的なシーンであるとみなし，時間の長い上位 10 シーンからフレーム画像を抽出し，それらを比較することによって映像間の類似度を算出する．

例えば，ホラーのような怖さを演出した映像には暗いシーンが多く，子供向けアニメなどの楽しい映像にはカラフルなシーンが多い傾向があると考えられる．そこで，代表シーンの色に基づいて類似度を算出することを考える．具体的には，画像の L*a*b*表色系を用いた色ヒストグラムを求め，これを比較することにより映像間の類似度を算出する．

映像v_mとv_nの代表シーンに基づく類似度$R_r(v_m, v_n)$の算出式を，式(1)に示す．

$$R_r(v_m, v_n) = \frac{1}{2}\sum_{i=1}^{|I^m|}\max_j R_h(I_i^m, I_j^n) + \frac{1}{2}\sum_{i=1}^{|I^n|}\max_j R_h(I_i^n, I_j^m) \quad (1)$$

ただし，I^m, I^nはそれぞれv_m, v_nから抽出された画像の集合であり，$R_h(I_i^m, I_j^n)$はI^mのi番目の画像とI^nのj番目の画像の色ヒストグラムの histogram intersection に基づく類似度である．

3.2 構造的特徴に基づく類似度算出

嗜好に影響を与える要因の一つとして，映像の雰囲気がある．提案手法では，映像の構造的特徴を利用して，これを捉えることを試みる．例えば，短時間に多くのカットが含まれていれば視聴者はスピード感を感じやすいであろう．そこで， このような映像内に生起するイベントの構造，すなわち，イベントが映像中のどの位置に，どの程度生起しているかを求め，これらを比較することにより映像の類似度を算出する．

†日本電信電話株式会社 NTT サービスエボリューション研究所

(注 1) Wikipedia, http://www.wikipedia.org/

提案手法では，イベントとしてカット点[5]，音楽区間[6]，発話区間[6]，テロップ区間[7]を用いる．映像全体を 30 区間に均等分割し，各区間におけるイベントの生起頻度を各次元の値とする構造的特徴ベクトルを作成する．イベントの正規頻度は，カットの場合は各区間におけるカット頻度（回/分）を，それ以外の場合は各区間長に対するイベント時間の割合を用いる．構造的特徴ベクトルを比較することにより，映像間の類似度を算出する．

映像v_mとv_nの構造的特徴に基づく類似度$R_s(v_m, v_n)$の算出式を，式(2)に示す．

$$R_s(v_m, v_n) = \sum_i k_i R_v(E_i^m, E_i^n) \qquad (2)$$

$$R_v(E_i^m, E_i^n) = \mathrm{e}^{-\|E_i^m - E_i^n\|} \qquad (3)$$

ただし，E_i^m, E_i^nはそれぞれ，v_m, v_nについて作成した，カット点，音楽区間，発話区間，テロップ区間のうちi番目のイベントの構造的特徴ベクトルであり，$\|E_i^m - E_i^n\|$はE_i^mとE_i^nの L2 ノルム距離である．また，k_iはi番目のイベントの構造的特徴に対する重みを決定するパラメータである．

3.3 メタデータと映像特徴に基づく推薦映像の決定

特徴量の計算量を削減するため，メタデータに基づく推薦リストの上位 100 件に限定して映像特徴に基づく類似度を算出する．最終的なアイテムの推薦スコアは，メタデータ，代表シーン，構造的特徴それぞれに基づく類似度の重み付き線形和により算出する．評価実験では，実験データを用いて学習し，パラメータを決定した．

4．評価実験

本手法の有効性を検証するため，メタデータのみに基づく推薦手法[2]と，提案手法の精度評価実験を行った．データセットとして，動画共有サイトにおける 16,188 本の映像を利用した．各映像には 1 つ以上のタグおよび 1 つ以上のコメントが付与されている．本実験ではコメント履歴をタイムスタンプ順にならべたものを視聴履歴として利用し，6 本以上の映像を視聴したユーザ 3,135 名を推薦対象とした．手法の精度は，各ユーザの 1〜5 本目の映像に基づいて推薦リストを作成し，6 本目を正解映像としたときの，上位 100 件の平均適合率（MAP）によって評価した．

実験結果を図 1 に示す．メタデータに基づく推薦手法では，比較するタグ数が少ない場合には類似度を正確に算出することが難しい．そのため，メタデータに基づく推薦手法は，付与されたタグ数が 1〜5 のとき，MAP が 0.0512 と，タグ数が多い場合に比べて低かった．一方，映像特徴を組み合わせた提案手法は，タグ数が 1〜5 のときの MAP が 0.0704 であり，メタデータのみに基づく推薦手法の精度を改善できた．タグ数別に見ると，タグ数が少ない場合ほど提案手法の効果があり，タグ数が 6 以上の場合の推薦精度についても，メタデータのみの手法と比較して精度が向上した．提案手法により，タグ数に依らず，雰囲気や編集のされ方といった特徴が類似している映像を効果的に推薦できることが期待される．

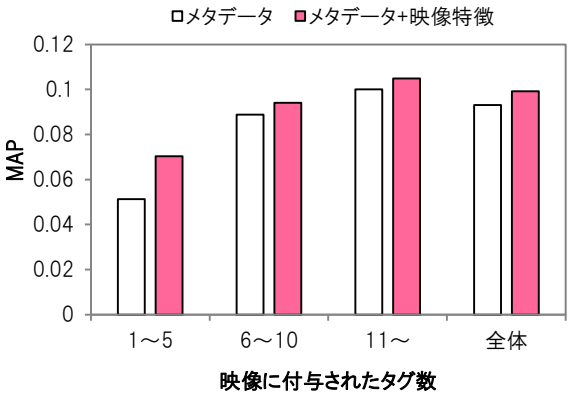

図 1　実験結果

5．まとめ

本研究では，メタデータに基づく推薦手法の精度向上を目指し，メタデータに加え，映像特徴に基づく手法を組み合わせた推薦手法を提案した．提案手法では，映像の代表シーンにおける画像の色ヒストグラムおよびカット，音楽，発話，テロップの構造的特徴に基づいて映像間の類似度を算出し，メタデータに基づく推薦結果と統合することにより，推薦するアイテムを決定する．提案手法の有効性を検証するため，動画共有サイトの映像を利用した実験を行った．実験の結果，提案手法がメタデータに基づく手法の推薦精度を改善できることを確認した．今後は，異なる映像データセットを利用した実験を行い，提案手法の有効性を検証する予定である．

参考文献

[1] D. Maltz and K. Ehrlich. Pointing the Way: Active Collaborative Filtering. In Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 202–209, 1995.

[2] 吉田 大我, 入江 豪, 佐藤 隆, 小島 明. タグランキングに基づく映像推薦. 第 10 回情報科学技術フォーラム, No. 2, pp. 1–6, 2011.

[3] Y. Song, J. Cao, Z. Chen, Y. Zhang and J. Li. Tag Transformer. In Proc. ACM International Conference on Multimedia (ACM MM), pp. 639–642, 2010.

[4] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang and M. Li. Online Video Recommendation Based on Multimodal Fusion and Relevance Feedback. In Proc. ACM International Conference on Image and Video Retrieval (CIVR), pp. 73–80, 2007.

[5] 谷口 行信, 外村 佳伸, 浜田 洋. 映像ショット切換え検出法とその映像アクセスインタフェースへの応用. 信学論, Vol. J79-D-II, No. 4, pp. 538–546, 1996.

[6] K. Minami, A. Akutsu, H. Hamada and Y. Tonomura. Video Handling with Music and Speech Detection. IEEE Multimedia, Vol. 5, No. 3, pp. 17–25, 1998.

[7] 桑野 秀豪, 倉掛 正治, 小高 和己. 映像データ検索のためのテロップ文字抽出法. 信学技報 PRMU, Vol. 96, No. 385, pp.39–46, 1996.