コンテンツの多様性を考慮したクロスドメイン推薦

富士谷 康† 村尾 和哉†† 望月 祐洋††† 西尾 信彦††

† 立命館大学大学院 情報理工学研究科 〒 525–8577 滋賀県草津市野路東 1 丁目 1–1
†† 立命館大学 情報理工学部 〒 525–8577 滋賀県草津市野路東 1 丁目 1–1
††† 立命館大学 総合科学技術研究機構 〒 525–8577 滋賀県草津市野路東 1 丁目 1–1
E-mail: †,† † †{fujiya, moma}@ubi.cs.ritsumei.ac.jp, ††{murao, nishio}@cs.ritsumei.ac.jp

あらまし ユーザの情報を獲得するドメインと推薦対象のドメインが異なるクロスドメインの推薦が研究されている．本研究では，コンテンツの多様性を持つドメインとしてテレビ番組を対象に，放送ごとの番組に適した書籍推薦を目指す．放送内容は多様なため，関連書籍の抽出に有効な特徴量は番組ごとに異なり，タイトルやキーワードといった語が有効な番組もあれば，抽象的な話題（トピック）が有効な番組もある．さらに，関連書籍が存在しない番組もある．このような多様性を考慮して，提案手法では TF-IDF と LDA の特徴量を併用し，書籍を抽出する．また，識別器を用いて推薦に不適な番組の判別および適切な特徴量の重みの決定を行う．評価では，提案手法が TF-IDF のみを用いた手法よりも推薦精度が高いことを確認し，推薦に不適な番組の判別と特徴量の重みの決定により推薦精度が向上した．

キーワード クロスドメイン推薦，テレビ番組，書籍，コンテンツベースフィルタリング

1. はじめに

ユーザの嗜好や状況に適した情報を抽出し，提示する推薦システムは，EC（Electronic Commerce）サイトの Amazon や，動画配信サービスの Netflix など，さまざまなサービスや製品に利用されており，売り上げの促進に貢献している．推薦システムを対象とした研究の中でも，クロスドメイン環境における推薦が注目されている．クロスドメインとは，推薦対象のアイテムが属するドメイン（ターゲットドメイン）と，ユーザについての知識を獲得するドメイン（ソースドメイン）が異なることをさす．ドメインの定義は研究者によってさまざまであるが [1]，ドメインをまたがった推薦が可能となれば，商品の併売やターゲットドメインでの利用履歴が十分に存在しないユーザへの推薦が期待できる．

推薦システムは主に，多数のユーザの履歴を活用する協調フィルタリングをベースとした手法と，アイテムに含まれる説明文などの属性を利用するコンテンツベースフィルタリングの 2 つに分けられる．クロスドメイン推薦において，協調フィルタリングの手法では，ドメインをまたがって利用するユーザの履歴の獲得が難しく，新規アイテムや新規ユーザへの推薦が困難であるコールドスタート問題がある．一方，コンテンツベースの手法では，ドメインをまたがったアイテム間の類似性を算出する必要があり，そのための特徴量の獲得が重要である．

本研究では，コンテンツの多様性を考慮したクロスドメイン推薦手法を提案する．コンテンツとは，アイテムに含まれる情報（内容）をさし，コンテンツの多様性をもつドメインのひとつとして，テレビ番組に着目する．ここでの多様性とは，番組が対象とする話題やジャンル，コーナ，シリーズなどをさす．推薦対象は書籍とし，放送ごとの番組に適した書籍の推薦に取り組む．コンテンツの多様性は書籍にも存在しており，書籍の内容やジャンルもさまざまである．本来は，ソースドメインおよびターゲットドメインの両方の多様性について考慮するべきであるが，本研究では，ソースドメインの多様性を考慮することから取り組む．

書籍を推薦対象としたのは，書籍と番組は関連があると考えたためである．ニュースや映画，料理番組など番組の放送内容はさまざまであり，それぞれの話題や番組に関連する書籍は多数存在する．すでに，書籍を含め，番組で紹介されたアイテムを提示するサービス (注1) も存在する．また，番組は書籍よりも手軽に見ることができ，番組の視聴者は，放送された話題に対する興味が喚起されると考えられる．例えば，ドラマやアニメでは原作本や出演者に関する書籍，経済ニュース番組ではビジネス書籍というように，番組に関連する書籍を視聴者に提示できれば，書籍の販売促進や書籍ストアの利用促進が期待でき，書籍の購入履歴やストアの利用履歴がないユーザに対しても推薦が可能になる．さらに，本研究の対象ではないが，番組に対するユーザの好みを獲得できれば，ユーザの嗜好を考慮した書籍などの推薦への展開も期待できる．

番組と書籍の類似性を算出するには，さまざまな種類の番組が存在することを考慮する必要がある．例えば，ニュース番組では事象を表すキーワード，トーク番組やアニメでは番組の出演者やタイトルなど単語そのものが有効である一方で，旅番組や時代劇では，特定の語より抽象的な概念であるトピックが有効である．そのため，複数の特徴量を用い，番組ごとに重視する特徴量を変えて書籍との類似性を算出できれば効果的な推薦に繋がる．さらに，クロスドメイン推薦の性質上，ターゲットドメインにソースドメインと関連するアイテムが存在しないことがある．例えば，通販番組や短時間の番組などでは，関連する書籍が存在しない場合がある．本研究では，このような番組を推薦に不適な番組と呼ぶ．推薦に不適な番組を判別し，推薦

(注1)：価格.com テレビ紹介情報（http://kakaku.com/tv/）など．

結果から省くことで，推薦精度の向上が期待できる．また，視聴履歴からユーザの嗜好を獲得する際にも，推薦に不適な番組の判別は利用しないといった方法が有効であると考えている．

本研究ではテレビ番組を対象としたが，アイテムごとに有効な特徴量が異なる点や，対応するアイテムが存在しない点は，雑誌やラジオなど多様なコンテンツを扱う他ドメインでも存在するため，提案手法は他ドメインでも有用であると考えている．

提案手法は，上述したコンテンツの特性を考慮するため，放送回ごとに作成される詳細な番組履歴を利用し，コンテンツベースの手法を採用する．TF-IDF と Latent Dirichlet Allocation（LDA）[2] の 2 つの特徴量によって，異なる抽象度で番組と書籍を表現し，それらを重み付けして併用することにより，番組に適した書籍の順位付けを行う．また，識別器を用いて，推薦を行う前に推薦に不適な番組の判別とともに番組ごとに特徴量の適切な重みの決定を行う．

本研究の構成は，次のとおりである．2 章で本研究に関連する研究について述べる．3 章で提案手法について述べ，4 章では，その評価実験について述べる．5 章で本研究をまとめる．

2. 関 連 研 究

本章では，クロスドメイン推薦およびテレビ番組推薦に関連する研究を紹介する．

2.1 協調フィルタリングを用いたクロスドメイン推薦

クロスドメイン推薦の研究は，コールドスタート問題の緩和や，他ドメインの情報を用いた精度向上，ユーザモデリング，多様性向上などを目的として行われている [1]．クロスドメイン推薦の研究では，ドメインに関する知識を必要としない協調フィルタリングの手法 [3], [4] をベースに行われていることが多い．

Li ら [5] は，評価を付けたアイテム数が全体の数に対して少ないと推薦の質が低下する sparsity 問題に対し，ソースドメインから獲得したアイテムに対する評価値パターンを，ターゲットドメインに転移し利用する手法を提案している．

中辻 [6] らは，ソーシャルネットワーク上でのユーザの関係から，ユーザをノード，ユーザ間の類似度を重みとしたエッジをもつグラフを構築し，ランダムウォークを用いてユーザ間類似度を計算し，推薦を行っている．

協調フィルタリングをベースとした手法は，ユーザの履歴を多く必要とする．しかし，サービスやシステムの開始当初には，多数の履歴の獲得は難しい．また，協調フィルタリングは，アイテムの内容や特性を考慮できないため，放送ごとの番組に適した書籍の抽出が困難になると考えられる．さらに，テレビ番組は逐次新たなものが放送され，放送時のみ視聴される番組も多いことから，協調フィルタリングをベースとした場合，コールドスタート問題が顕著になることが予想される．

2.2 コンテンツベースフィルタリングを用いたクロスドメイン推薦

Fukazawa ら [7] は，ユーザの実世界での行動（タスク）と，それに関連する検索語に着目して構築したモデルを利用し，ニュース，テレビ番組，モバイルサイトに対するユーザ個人の評価値を用いて，嗜好に適した推薦を行っている．アイテムとユーザの嗜好をタスクで表現することで，ドメインをまたがった推薦を可能にしている．しかし，タスクに抽象化すると，個々のアイテムの特徴が失われてしまう．例えば，異なるドラマを見た場合でも，タイトルは考慮されず「ドラマを見る」というタスクに抽象化されるため，多様なコンテンツを表現できない．

Fernández-Tobías ら [8] は，異なるドメインに属するアイテム間の類似度を，Wikipedia の記事間のリンク関係から獲得し，興味のある場所に適した音楽の推薦を実現している．しかし，この手法では，Wikipedia に記載されていないアイテムの推薦ができない．さらに，記事として記載されている内容が，アイテムの特徴を十分に表しているとは限らない．例えば，テレビ番組の記事には，放送局や出演者といった概要だけが記載され，実際の放送内容は記述されていないことが多い．

Elkahky ら [9] は，多数のユーザの検索クエリとアイテムの説明文，およびクリックストリームデータ（ページ遷移情報）を用いてアイテムとユーザの嗜好を表すことで，Windows アプリケーション，ニュース，映画・テレビを対象に推薦を行っている．しかし，この手法では，多くのユーザの履歴が複数のドメインで必要であり，検索エンジンやポータルサイトなど，多数の履歴を保持するごく一部のサイト運営者しか利用できない．

2.3 テレビ番組推薦

日常的にユーザが視聴するテレビには，ユーザの嗜好が現れ，視聴履歴は，嗜好が反映されたライフログと考えることができる．このような情報からユーザに適した情報提示ができれば有益である．この視聴履歴を用いた番組推薦の研究が行われている [10] [11] [12]．番組の情報を利用し，ユーザや番組に適した情報を抽出する点では本研究と共通しているが，これらの研究の推薦対象は番組であり，番組以外のアイテムを推薦するクロスドメインの推薦やその評価は行われていない．

3. 提 案 手 法

本研究では，多様なコンテンツをもつ番組に適した書籍を抽出する手法を提案する．前述のコールドスタート問題を回避するために，コンテンツベースの手法を採用する．

提案手法は，コンテンツの多様性に対応するために，TF-IDF と LDA を特徴量として用いる．TF-IDF は単語に対し重み付けする手法であるが，同形異義語が存在していたり単語への重み付けが適切に行えないと，関連が薄い書籍が抽出されることがある．一方，LDA は説明文などの背景に存在するトピックを捉える手法で，番組や書籍の話題を考慮できる．しかし，説明文が十分になかったり書籍や番組から推薦に有効なトピックを得られないと関連が薄い書籍が抽出されることがある．そこで，これらの特徴量を併用し補うことで，書籍を抽出する．それぞれの特徴量について番組と書籍の距離（非類似度）を算出し，書籍を順位付けした後，2 つの順位を足し合わせることで，番組に対して書籍を順位付けする．TF-IDF と LDA のそれぞれの順位を足し合わせる際に，両者を考慮する割合として重みを導入し，これを本研究では利用比率と呼ぶ．TF-IDF の利用比率を高めることで，個々の単語を重視する．一方，TF-IDF の利用比率を下げ，LDA の比率を高めるとトピックを重視し

た推薦を行う．

加えて，推薦に不適な番組の識別と番組に適した特徴量の利用比率を定めるため，番組ごとに複数の利用比率で抽出した書籍に対して，事前に与えた評価値を利用する．この評価値は，書籍ごとに番組と関連があるかという観点で，少数の分析者が与えることを想定している．これを利用すると，番組ごとに推薦に不適か否かを判断でき，推薦に適している場合に適切な利用比率を決定できる．

本章ではまず，本研究で用いる番組情報および書籍情報について述べる．続いて，特徴量の抽出とそれらを用いた推薦手法について述べる．最後に，推薦に不適な番組の判別と利用比率を定めるための識別器の構築について述べる．

3.1 番組情報と書籍情報

番組情報は，放送内容が EPG や Wikipedia などよりも詳細に記述されているものを扱う．この番組情報は，1 章で述べた，番組で紹介されたアイテムを提示するサービスで利用されており，人力で作成され，放送終了直後に公開される．この属性として，タイトル，放送局，番組開始時刻や終了時刻，出演者やジャンル，コーナがある．コーナは，コーナ名と複数のサブコーナを含み，サブコーナは，サブコーナ名，開始時刻と終了時刻，説明文，キーワードを含む．本研究では，サブコーナの説明文をまとめて番組説明文と呼ぶ．このような，番組に関する詳細な情報を用いて，多様な放送内容を考慮し，放送ごとの番組に適した書籍の推薦を実現する．

書籍情報の属性には，書籍 ID，タイトルおよびシリーズ名，書籍説明文，著者，出版社，ジャンル，価格，販売開始日などがある．推薦対象の書籍は，アダルトなど一部ジャンルを省いた約 20 万冊とした．

3.2 特徴量の抽出

TF-IDF によるベクトルと LDA を用いたトピックベクトルの 2 つのベクトルで，番組と書籍の特徴を表現する．異なる抽象度の特徴量を組み合わせることで，推薦精度を向上させることを目指す．

3.2.1 特徴語の獲得

まずはじめに，番組および書籍の情報から，特徴語を獲得する必要がある．番組情報からは，タイトル，コーナ名，サブコーナ名，説明文，キーワード，出演者を利用する．書籍情報では，タイトル，シリーズ名，著者，説明文を利用する．それぞれの文章から，MeCab(注2) を用いて形態素に分け，単語を得る．今回は，形態素のうち数詞や非自立語などを除いた名詞を用い，一部の語については，ストップワードとして除去した．また，形態素解析器には書籍タイトルやシリーズ名，著者から構築した辞書を登録した．本研究では複合名詞を考慮するために，特徴語として単語ユニグラムおよび単語バイグラムを合わせて利用した．

3.2.2 TF-IDF を用いた特徴量抽出

テキストマイニングで一般的に用いられる TF-IDF 手法を利用し，3.2.1 節で獲得した特徴語からベクトルを構築する．まず，書籍に関して述べる．書籍 b に含まれる特徴語 w の重み $t_{w,b}$ を式（1）で示すように計算する．$tf_{w,b}$ は書籍 b における特徴語 w の出現回数 $freq(w,b)$ の対数であり，式（2）で得られる．idf_w^{book} は IDF（逆文書頻度）と呼ばれ，書籍総数 N^{book} および，特徴語 w が出現する書籍数 df_w^{book} を用いて．式（3）から得られる．係数 $n_{w,b}^{book}$ は文章長によって正規化するために用いられ，式（4）で得られる．

$$t_{w,b} = \frac{tf_{w,b} \cdot idf_w^{book}}{n_{w,b}^{book}} \tag{1}$$

$$tf_{w,b} = \log\left[freq(w,b)+1\right] \tag{2}$$

$$idf_w^{book} = \log\left(\frac{N^{book}+1}{df_w^{book}+1}\right) \tag{3}$$

$$n_{w,b}^{book} = \sqrt{\sum_{w \in V}\left(tf_{w,b} \cdot idf_w^{book}\right)^2} \tag{4}$$

番組でも同様に，特徴語から TF-IDF ベクトルを獲得する．番組 p に含まれる特徴語 w の重み $t_{w,p}$ を $tf_{w,p}$（式（6））, idf_w^{tv}（式（7））および $n_{w,p}^{tv}$（式（8））をもとに式（5）で示すように計算する．ここで，出現した特徴語の IDF 値（式（7））が必要になるが，今回は利用する情報数が多いドメインである書籍情報から構築したもの（式（3））を用いた．特徴語の IDF 値はドメイン間で異なるが，この考慮については，今後の課題とする．

$$t_{w,p} = \frac{tf_{w,p} \cdot idf_w^{tv}}{n_{w,p}^{tv}} \tag{5}$$

$$tf_{w,p} = \log\left[freq(w,p)+1\right] \tag{6}$$

$$idf_w^{tv} = idf^{book} = \log\left(\frac{N^{book}+1}{df_w^{book}+1}\right) \tag{7}$$

$$n_{w,p}^{tv} = \sqrt{\sum_{w \in V}\left(tf_{w,p} \cdot idf_w^{tv}\right)^2} \tag{8}$$

3.2.3 LDA を用いた特徴量抽出

3.2.1 節で獲得した特徴語をもとに，LDA を用いてトピックベクトルを構築する．LDA は，文書に出現する単語とその背景に隠れて存在するトピック（話題）の関係を確率的に表したトピックモデルのひとつであり，テキストマイニングを始めとして協調フィルタリングなど幅広い分野で利用される．LDA では，特徴語 $w(w \in W)$ の列で表された文書群 $d(d \in D)$ とトピック数 K を入力することで，トピック $z_n(n = 1, ..., K)$ における語 w の確率分布 $P(w|z_n)$ および，各文書 d におけるトピック z_n の確率分布 $P(z_n|d)$ を推定する．本研究では，ひとつの文書（書籍または番組）のトピックの分布 $P(z_n|d)$ を用いて，式（9）で表したベクトル v_d をトピックベクトルと呼ぶ．トピックベクトルの次元数はトピック数 K であり，要素はトピックの強さを表す．

$$v_d = (P(z_1,d), P(z_2,d), ..., P(z_K,d)) \tag{9}$$

LDA の実装の一つである Mallet(注3) を用いて，まずは，書籍群から $P(w|z_n)$ と $P(z_n|b)$ を獲得し，書籍 b のトピックベク

(注2)：http://taku910.github.io/mecab/

(注3)：http://mallet.cs.umass.edu/

表 1　番組に対する推薦書籍の例

利用比率	順位	$rank_{\mathrm{tfidf}}$	$rank_{\mathrm{topic}}$	書籍タイトル	関連語
$\alpha = 0$	1	60	1	中国 大気汚染　PM2.5 の霧，どうなる日本への影響	汚染 物質 濃度
	2	32	2	電線一本で世界を救う	汚染物質 物質 汚染
	3	36	3	エコで世界を元気にする！	プラスチック 汚染 生産
$\alpha = 0.5$	1	6	4	複合汚染（新潮文庫）	有害物質 物質 汚染
	2	2	16	有害化学物質の話	プラスチック 有害 物質
	3	7	11	循環型社会　持続可能な未来への経済学	リサイクル 有害 ゴミ
$\alpha = 1$	1	1	26	プラスチック材料活用事典	プラスチック リサイクル 種類
	2	2	16	有害化学物質の話	プラスチック 有害 物質
	3	3	33	図解入門よくわかる最新 プラスチックの仕組みとはたらき	プラスチック 種類 素材

トル v_b を得る．トピック数 K は本研究では 100 とした．

続いて，番組のトピックベクトルを獲得する．番組は書籍に比べ，多様な話題が放送されることが多い．例えば，朝の情報番組では，ニュースや特集，天気予報などさまざまな話題があり，ニュースにおいても，殺人事件や交通事故など多様な話題が放送される．そのため，番組群からトピックを構築すると，複数の話題が混在したトピックが生成され，書籍との類似性を算出することが困難になることが予想される．そのため，書籍群から構築したトピックを用いて番組を表す．書籍群の $P(w|z_n)$ を用いて，ギブスサンプリングによって，番組 p の $P(z_n|p)$ を獲得し (注4)，トピックベクトル v_p を得る．番組のトピックベクトル v_p の次元数は，書籍と同じく $K = 100$ であり，要素は，番組における書籍群から構築したトピックの強さを表す．

3.3　TF-IDF と LDA を併用した推薦

TF-IDF で得られた特徴語の重みベクトルと LDA によって獲得したトピックベクトルの 2 つの特徴量をもとに，番組と書籍の距離（非類似度）を算出する．はじめに，TF-IDF の重みベクトルを用いて番組ごとにすべての書籍と番組の距離を算出し，昇順に並べ替え，順位 $rank_{\mathrm{tfidf}}$ を算出し，上位 k 冊を抽出する．前もって，TF-IDF の上位書籍を抽出することで，書籍と関連が薄い書籍にもかかわらず LDA のトピックベクトルと近くなり，誤って抽出されることを防ぐ．距離 $distance$ はコサイン類似度を変換した式（10）を用い，$\boldsymbol{p}$ は番組，$\boldsymbol{b}$ は書籍の特徴量を表す．k を大きくすると，トピックベクトルを重視できるようになるが，本研究では $k = 100$ とした．

$$distance\,(\boldsymbol{p}, \boldsymbol{b}) = \frac{2 \cdot cos^{-1}\,(cos\,(\theta))}{\pi} \tag{10}$$

$$cos\,(\theta) = \frac{\boldsymbol{p} \cdot \boldsymbol{b}}{\| \boldsymbol{p} \| \| \boldsymbol{b} \|} \tag{11}$$

次に，先ほど抽出した上位 k 冊の書籍に対して，トピックベクトルによって距離を算出し，昇順に並び替え，順位 $rank_{\mathrm{topic}}$ を得る．番組 p に対する書籍 b のスコア s を，TF-IDF の順位 $rank_{\mathrm{tfidf}}$ と，トピックによる順位 $rank_{\mathrm{topic}}$，およびそれらの利用比率 α から，式（12）を用いて算出する．スコア s をもとに，昇順に並び替えることで，番組に対する書籍の順位付けを行う．もし，複数の書籍が同一スコアであった場合には，TF-IDF による距離と，LDA による距離を掛けたものが小さい順に上位に順位付けする．

$$\begin{aligned} s_{p,b} =& \alpha \cdot rank_{\mathrm{tfidf}}\,(p, b) \\ &+ (1 - \alpha) \cdot rank_{\mathrm{topic}}\,(p, b) \end{aligned} \tag{12}$$

$\alpha = 1$ のとき，推薦書籍は TF-IDF のみでの順位となり，$\alpha = 0$ のとき，TF-IDF での上位 k 冊をトピックベクトルで順位づけたものとなる．$\alpha = 0.5$ のとき，TF-IDF とトピックベクトルの順位を同等に考慮する．TF-IDF とトピックベクトルを併用することで，TF-IDF で人の名前といった同形異義語などの不適切な語により抽出された書籍の順位を下げ，番組とトピックが類似する書籍の順位を高めることが期待できる．

3.4　推薦書籍の例

3.3 節で述べた手法で抽出した，推薦書籍の例について記す．2015 年 10 月 29 日放送のクローズアップ現代という番組は，「海に漂うマイクロプラスチックが海洋環境汚染に与える影響」についての特集であった．提案手法によって抽出した $\alpha = 0$, 0.5，1 に対する推薦書籍を表 1 に示す．表中の関連語は，番組と書籍の特徴語において TF-IDF 値の積が大きい 3 語を示す．$\alpha = 1$ では，TF-IDF が重視されるため，番組の説明文に頻出した「プラスチック」という具体的な語に関連する書籍が抽出されており，図鑑などの書籍が抽出されている．一方で，$\alpha = 0$ では LDA によるトピックベクトルが重視されるため，番組のトピックに適した，環境問題や大気汚染に関連する書籍が抽出される．それぞれの順位を半分ずつとした $\alpha = 0.5$ についても同様に，環境問題に関わる書籍が抽出される．

3.5　推薦に不適な番組の判別と TF-IDF と LDA の利用比率の決定

クロスドメイン推薦の特性上，ソースドメインのアイテムに対応するターゲットドメインのアイテムが存在しない場合がある．これにより例えば，通販番組や短時間の番組にはそもそも関連する書籍が存在しないのにもかかわらず，推薦手法によって関連性が薄い書籍が抽出されてしまうという問題が生じる．このような，推薦に不適な番組を判別し，推薦結果から省くことで，推薦精度の向上が期待できる．また，TF-IDF とトピックベクトルの利用比率を表す α は，番組によって適切な値が異なると考える．例えば，バラエティやトーク番組では出演者，アニメや映画ではタイトルといった語そのものが有効である．一方で，料理番組や時代劇では，説明文の背景に存在するトピックも有効であると考えられる．

(注4)：Mallet に含まれる機能である TopicInferencer を利用した．

推薦に不適な番組の判別と，利用比率 α を定めるために識別器を用いる．あらかじめ，一定期間の番組に対し，複数の利用比率 α で書籍を抽出し，それぞれの書籍に対して評価値を付けたデータを準備する．評価値は，少人数の分析者により番組と書籍が関連があるかという観点に基づいて付与するか，A/B テストによりランダムに定めた利用比率を用いて多数の被験者に提示しどの比率が最もページ遷移や購買を起こしたかを測ることでの獲得を想定している．また，本研究では利用比率 α は 0，0.5，1 の 3 つのうちのいずれかを用いる．

続いて，番組 p のそれぞれの利用比率で得られた評価値を正解情報として利用比率 $\alpha = a$ の精度 $Prec_{p,\alpha=a}$ を求める．この精度は，上位 k 冊の適合率を表す Precision@k を算出することで得る．ベクトル $Prec_p = (Prec_{p,\alpha=0}, Prec_{p,\alpha=0.5}, Prec_{p,\alpha=1})$ をもとに，k-means により番組をクラスタリングし，各クラスタにおける Precision@k の平均を算出する．このとき，すべての利用比率で精度が高いクラスタ（利用比率に関係なく正しく推薦が行えているクラスタ）や，すべての利用比率で精度が低いクラスタ（推薦に不適な番組のクラスタ），TF-IDF またはトピックベクトルのいずれかが有効なクラスタなどが現れると考えられる．この得られたクラスタに対して，その精度をもとに，分析者が推薦に不適なクラスタと定めるか，それ以外のクラスタでは利用する利用比率を定める．このクラスタを目的変数，番組の特徴量を説明変数として，識別器を学習する．番組の特徴量からクラスタを推測し，推薦に不適なクラスタに属するかを判別する．それ以外のクラスタに属する場合は，先に定めた属するクラスタの利用比率を利用する．本研究では，識別器は線形 SVM を用い，その実装の 1 つである liblinear(注5) を利用する．特徴量として，番組の TF-IDF ベクトルとトピックベクトルおよび番組のカテゴリを用いる．

4. 評価実験

提案手法を評価するため，2 つの実験を行った．1 つ目は，被験者が自身の興味とは無関係に多数の番組に対して評価値をつけ，TF-IDF と LDA によって獲得したトピックベクトルを併用した推薦の有効性を確認する．また，被験者による評価値をもとに，番組を k-means によってクラスタリングし，クラスタごとに推薦に不適な番組や適切な利用比率を定める．さらに，識別器を用いて，番組が属するクラスタを判別できるかを評価し，最後に，定めた利用比率を用いることで精度が向上するかを評価する．2 つ目は，被験者自身に番組を選択してもらい，それらに対し書籍を抽出する．番組ごとに推薦された書籍が被験者にとって興味があったかを評価する．

4.1 実験 1：多数番組に対する評価

少数の被験者を対象に，多数の番組に関する評価値を付けてもらい，この評価値を用いて，提案手法を評価する．

4.1.1 評価環境

被験者は，20 代から 50 代の男性 5 名である．1 日間（2015 年 10 月 14 日）に東京地区で放送された，207 番組を対象とする．提案手法の利用比率を $\alpha = 0$，0.5，1 の 3 種類で番組ごとに書籍を順位づけし，上位 5 冊ずつ，計 15 冊を被験者に提示する．被験者には，提示された書籍それぞれに「提示された書籍は番組と関連があるか（良い推薦書籍か）」という観点で，2 段階での評価値（「関連がある」と「関連がない」）をつけてもらい，提案手法による書籍推薦の精度を算出する．

(注5)：https://www.csie.ntu.edu.tw/~cjlin/liblinear/

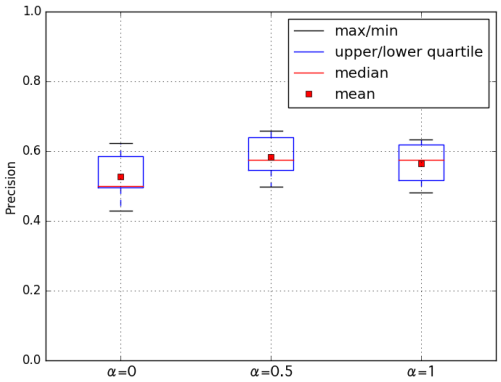

図 1　各被験者の Precision@5（実験 1）

4.1.2 評価結果

評価尺度は，ある番組に対し，被験者に提示した上位 5 冊のうち「関連がある」と評価が付けられた書籍の割合を表す Precision@5 と，「関連がある」と評価付けられた各書籍での順位における Precision@k を算出し上位 5 冊の平均を算出した Average Precision を利用する．それぞれの被験者における番組ごとの Precision@5 の平均を図 1 に，Average Precision を図 2 に示す．

Precision@5 および Average Precision のいずれでも，$\alpha = 0.5$（TF-IDF と LDA の順位を半分ずつとしてスコアを算出）の結果は，$\alpha = 1$（TF-IDF のみ）の結果と比べて，精度が向上している．これは，LDA を用いてトピックを考慮することで，適切でない書籍の順位が下がったためだと考えられる．一方で，$\alpha = 0$ としてトピックを優先した場合には，$\alpha = 1$ と比べて精度が低下している．これは，多くの番組において，抽象度が高い話題によって関連が薄い書籍が抽出されてしまったことが原因だと考えられる．

番組の Precision@5 および Average Precision について，$\alpha = 0.5$ が $\alpha = 1$ よりも有意に高いかについて，有意水準を 5 %として t 検定を行った．Precision@5 での p 値は 0.0212，Average Precision での p 値は 0.0358 となり，それぞれ $\alpha = 0.5$ が有意に高いという結果になった．1 日間という一定期間内のすべての番組を対象とした場合，$\alpha = 0.5$ の手法が，$\alpha = 1$ よりも関連する書籍を提示できるといえる．

4.1.3 番組と特徴量の関係の分析

4.1.2 節の評価結果はすべての番組に対して同一の α の値を適用した場合のものである．そこで，番組ごとに有効な特徴量が異なるかを分析する．3.5 節で述べた方法により，それぞれの番組において，$\alpha = 0$，0.5，1 の各比率での番組 p の書籍推薦の精度 $Prec_{p,\alpha=a}$ を算出し，3 次元のベクトル $Prec_p$ を

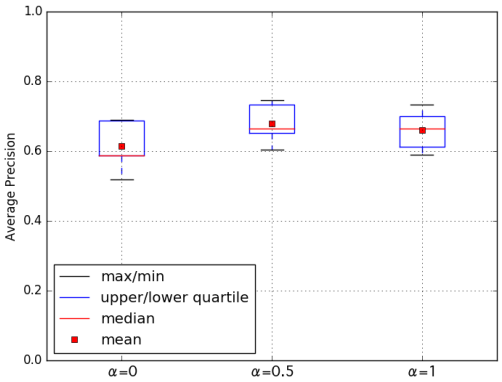

図 2　各被験者の Average Precision（実験 1）

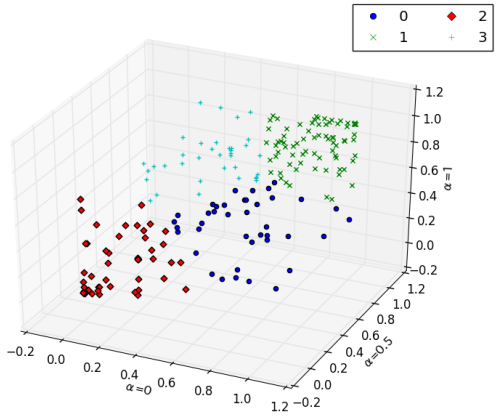

図 3　番組 p における精度のベクトル $Prec_p$ のクラスタリング結果

表 2　クラスタごとに定めた利用比率 α

クラスタ	利用比率 α
0	0
1	0.5
2	None（推薦に不適な番組）
3	1

k-means によってクラスタリングを行った．精度の指標として，番組における被験者の Precision@5 の平均を利用する．本研究では，$k = 4$ とし，番組を 4 つのクラスタに分けている．これは，番組は，すべての利用比率で推薦が有効な番組群，推薦に推薦に不適な番組群，TF-IDF が有効な番組群，トピックベクトルが有効な番組群の 4 つに分けることができると考えたためである．

クラスタリング結果を可視化させたものを図 3 に示す．印はクラスタのラベル（番号）を表しており，印の重なりを減らし見やすくするために，各比率での精度のベクトル $Prec_p$ の各要素にランダムな小さな数値（ノイズ）を加えて表示している．また，各クラスタの Precision@5 を図 4 に示す．図中のエラーバーは標準偏差を示す．図 4 に示す結果より，クラスタリングにより，クラスタ 0（40 番組）は $\alpha = 0$ で推薦精度が高い番組（LDA が有効である番組），クラスタ 1（77 番組）はいずれの比率でも推薦精度が高い番組，クラスタ 2（54 番組）はいずれの比率でも推薦精度が低い番組（推薦に不適な番組），クラスタ 3（36 番組）は $\alpha = 1$ で推薦精度が高い番組（TF-IDF が有効である番組）の 4 つに分かれた．この結果をもとに，クラスタごとに定めた利用比率を表 2 に示す．つまり，視聴者が視聴した番組が属すクラスタを判別し，適切な α を指定し，番組ごとに有効な特徴量を用いれば，精度を高めることができるといえる．

4.1.4　識別器の構築とその評価

番組ごとに適切な α の値を設定するため，3.5 節で述べた識別器を構築し，番組の特徴量から属するクラスタを判別する精度を 10-closs-validation によって評価した．クラスタの識別結果の混同行列を表 3 に示す．平均適合率は 0.5169 となった．クラスタ 1 とクラスタ 2 の適合率は一定程度あるが，クラスタ 0 とクラスタ 3 の適合率は低い．クラスタ 0 はクラスタ 2 に誤識別され，クラスタ 3 はクラスタ 1 に誤識別されることが多い．この識別精度の低さの原因は，話題の多様さに対して訓練データが少ないことが考えられる．実験では，1 日間の番組情報を対象としたため，放送内容が類似する番組が多くなかった．識別精度を向上させるためには，少なくとも 1 週間程度の番組に対する評価値を用いる必要があると考えている．

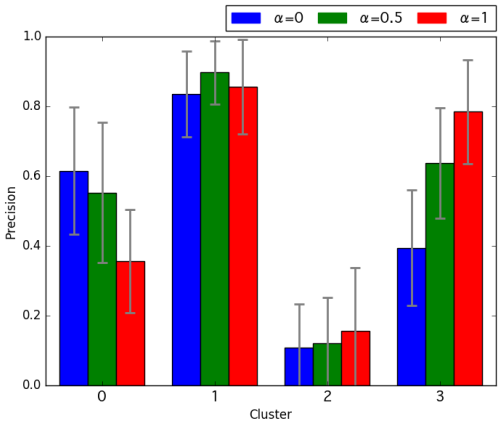

図 4　各クラスタの平均の Precision@5

表 3　クラスタの識別結果の混同行列

		推測したクラスタ				
		0	1	2	3	適合率
正解のクラスタ	0	6	13	21	0	0.1500
	1	3	52	13	9	0.6753
	2	4	9	40	1	0.7407
	3	0	21	6	9	0.2500

4.1.5　推薦に不適な番組の判別および利用比率の決定に関する評価

4.1.4 節で構築した識別器を用いて，番組の特徴量から番組が属するクラスタを推測し，そのクラスタに応じた表 2 の利用比率 α を用いて書籍を推薦する．本節では被験者の評価値をもとに，推薦された書籍が番組と「関連がある」とされた精度を計測する．結果を表 4 に示す．識別器を用いて推測した利用比率を用いた場合には被験者の Precision@5 の平均は 0.7128 となり，すべて $\alpha = 0.5$ とした時の 0.5832 よりも，約 0.22 改善

表 4　α を固定値，クラスタラベル，正解としたときの被験者の Precision@5 の平均

α の定め方	Precision@5
すべて $\alpha = 0.5$ とする	0.5832
識別器により判別したクラスタのラベルを利用する	0.7128
真の正解のクラスタのラベルを利用する	0.7969

した．これは，識別器を用いたことで，推薦に不適な番組を除くことができ，かつ番組ごとに適切な利用比率を定めたためであると考えられる．

4.2　実験 2：被験者選択番組に対する推薦結果の評価

被験者自身に番組を選択してもらい，提案手法によって推薦した書籍は，被験者が興味をもつものであるかを評価する．

4.2.1　評 価 環 境

被験者は，20 代から 50 代の男性 14 名，20 代女性 1 名である．2015 年 9 月 1 日から 10 月 31 日までの番組を対象に，被験者に 10 番組程度を選択してもらった．番組ごとに，提案手法（$\alpha = 0$，0.5，1）および，LDA のみの計 4 種類の手法の推薦結果を，被験者に提示する．提案手法の $\alpha = 0$ は LDA の順位を重視するが，TF-IDF によってあらかじめ上位 $k = 100$ 冊を抽出した後に，LDA の距離によって順位付けする．一方，LDA のみの手法は，すべての書籍を対象として LDA の距離によって順位付けする．手法ごとに 10 冊を提示し，それぞれの書籍に対して 5 段階評価で「推薦書籍に興味があるか」を尋ねた．また，比較対象として，2015 年 9 月の売り上げランキングの上位 10 冊の書籍を番組に関係なく被験者に提示し，同様に 5 段階評価で尋ねた．

4.2.2　評 価 結 果

被験者が選択した番組数の平均は，11.7 番組であった．5 段階評価での評価値において，4 以上を正解書籍とした場合の被験者ごとの平均の Precision@10 を図 5 に，Average Precision を図 6 に示す．さらに，5 段階評価の 4 以上を正解書籍とした，売り上げランキングの Precision@10 と Average Precision を図 7 に示す．

Precision@10 および Average Precision ともに被験者ごとの平均は $\alpha = 0.5$ で最も高くなった．Precision@10 において，$\alpha = 0.5$ は $\alpha = 1$（TF-IDF のみ）と比較して 0.1882 から 0.2144 に向上している．また，Precision@10 と Average Precision ともに LDA のみが最も低い．これはトピックだけを用いると，番組と関連が薄い書籍が抽出されてしまうためである．しかし，トピック単体では精度が低い LDA も，TF-IDF と組み合わせることで，精度が向上している．また，売り上げランキングは，漫画といった特定のジャンルに偏っているため，被験者によって精度のばらつきが大きく，中央値は提案手法（$\alpha = 0.5$）のそれよりも低い．このことから，提案手法は売り上げランキングの書籍と同等の精度で番組に関連した書籍が得られているといえる．

続いて，各被験者の Precision@10 および，Average Precision について，$\alpha = 0.5$ が $\alpha = 1$（TF-IDF のみ）よりも，優位に高いかについて，有意水準を 5 %として t 検定を行った．

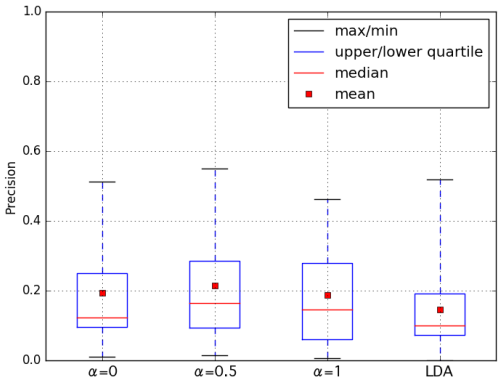

図 5　各被験者の Precision@10（実験 2）

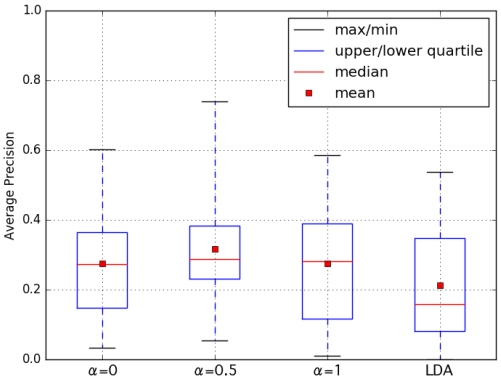

図 6　各被験者の Average Precision（実験 2）

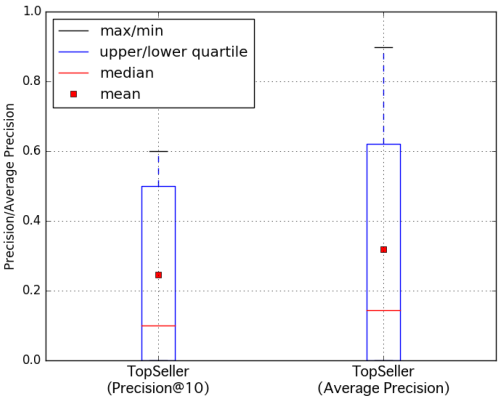

図 7　売り上げランキングに対する被験者の Precision@10 と Average Precision

Precision@10 において p 値は 0.0221，Mean Average Presicion において p 値は 0.0517 となった．Average Precision では，有意差は生じなかったものの Precision@10 では有意な差が生まれたことから，興味を喚起する推薦が可能であることを確認した．

4.2.3　推薦に不適な番組の判別および利用比率の決定に関する評価

3.5 節で述べた手法により構築した識別器を用いて，被験者

が選択した番組が属するクラスタを推定し，クラスタごとに定めた TF-IDF と LDA の利用比率を用いることで，推薦精度が向上するかを評価する．評価値は，実験 1 で得た 1 日間のすべての番組に対するものを利用する．クラスタごとの利用比率は 4.1.3 節で述べたものと同じく，表 2 に示したものを利用する．

被験者が選択した番組の総数は 176 番組であり，推定したクラスタごとの番組数は，クラスタ 0 が 15，クラスタ 1 が 77，クラスタ 2 が 75，クラスタ 3 が 9 となった．推薦精度は，すべての番組で利用比率を $\alpha = 0.5$ としたときの被験者の Precision@5 の平均が 0.2144 であった．一方，識別器を用いて利用比率を定めた場合には 0.2338 となった．有意な差があるかを調べるため，t 検定を行ったところ p 値は 0.4182 となった．識別器を用いても精度が向上したとは言えない結果となったが，これは，被験者は番組を 2ヶ月間に放送されたものから選択しているため，それらの番組を識別するのに十分な訓練データが存在しなかったことが原因であると考えられる．4.1.4 節で述べたように，1 日間の番組を対象にした場合には精度が向上していることから，訓練データを十分に増やすことにより，識別精度が高まることで，適切に推薦に不適な番組の判別とともに適切な利用比率を定めることができ，推薦精度は向上すると考えられる．

5. おわりに

多様なコンテンツをもつドメインを対象に，クロスドメイン推薦に有効な推薦手法を提案し，テレビ番組に関する情報を用いて，書籍を推薦するシステムを構築した．提案手法は，TF-IDF と LDA を適応的に併用し，それぞれの欠点を補い，有効な特徴量が番組ごとに異なることを考慮した．また，クロスドメインの性質上，ターゲットドメイン（書籍）にソースドメイン（番組）と関連があるアイテムが存在しない場合があり，このような推薦書籍が抽出できない推薦に不適な番組を識別器によって判別し，番組ごとに有効な特徴量を定めた．被験者による評価実験を行い，TF-IDF と LDA の併用が有効であることを示し，さらに，識別器を用いることで推薦精度が高まることを確認した．

今後の課題や展望として，ユーザの嗜好を考慮した推薦がある．既存の機械学習やユーザモデリングの手法を用いるなどして，視聴履歴などから視聴履歴から嗜好を抽出し，番組やその関連書籍に対するフィルタリングを行う．特に，テレビ番組を対象とした場合，被験者が番組のどの部分に興味をもって視聴しているかや，複数人の利用者が居る場合の視聴者の推定，ながら見やテレビをつけっぱなしにしているといった視聴形態の考慮も必要となる．

他にも，提案手法と協調フィルタリングのハイブリッドの手法の開発が挙げられる．コンテンツベースの手法では，書籍の販売実績や人気度合いなど，書籍の質の考慮ができない．また，セレンディピティも協調フィルタリングに比べ，一般的に劣るとされている．そのため，不適な番組を除去するなどを行ったうえで，本手法による類似度の算出を用い，その結果を協調フィルタリングの入力とするなどして，推薦書籍を抽出する．

また，推薦対象のドメインとして，本研究では書籍を用いたが，番組に関連するアイテムは，商品では，音楽や映画，食品，家電などがあり，他にも旅行や飲食などのサービスがある．そのため，番組ドメインも含めて，ソースドメイン・ターゲットドメインを拡大することや，TF-IDF や LDA 以外の他の類似度の利用を検討する．

謝辞　書籍情報をご提供いただいたシャープ株式会社に，また，番組情報をご提供いただいた株式会社ワイヤーアクションに，深く感謝致します．

文　献

[1] Iván Cantador, Ignacio Fernández-Tobías, Shlomo Berkovsky, and Paolo Cremonesi. Cross-domain recommender systems. In *Recommender Systems Handbook*, pp. 919–959. Springer, 2015.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022, 2003.

[3] Pinata Winoto and Tiffany Tang. If you like the devil wears prada the book, will you also enjoy the devil wears prada the movie? a study of cross-domain recommendations. *New Generation Computing*, Vol. 26, No. 3, pp. 209–225, 2008.

[4] 柳原正, 帆足啓一郎, 小野智弘. クロスメディア型レコメンデーションの提案と評価. 日本データベース学会論文誌, Vol. 8, No. 2, pp. 13–18, 2009.

[5] Bin Li, Qiang Yang, and Xiangyang Xue. Can movies and books collaborate?: Cross-domain collaborative filtering for sparsity reduction. In *Proceedings of the 21st International Jont Conference on Artifical Intelligence, IJCAI'09*, pp. 2052–2057, 2009.

[6] 中辻真, 藤原靖宏, 内山俊郎. ユーザグラフ上のランダムウォークに基づくクロスドメイン推薦. 人工知能学会論文誌, Vol. 27, No. 5, pp. 296–307, 2012.

[7] Yusuke Fukazawa and Jun Ota. User-centered profile representation for recommendations across multiple content domains. *Int. J. Know.-Based Intell. Eng. Syst.*, Vol. 15, No. 1, pp. 1–14, 2011.

[8] Ignacio Fernández-Tobías, Iván Cantador, Marius Kaminskas, and Francesco Ricci. A generic semantic-based framework for cross-domain recommendation. In *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems, HetRec '11*, pp. 25–32. ACM, 2011.

[9] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, pp. 278–288, 2015.

[10] 山田一郎, 宮崎勝, 住吉英樹, 古宮弘智, 田中英輝. ランダムウォークを利用した番組類似性評価. 情報処理学会研究報告, Vol. 2012, No. 12, pp. 1–7, 2012.

[11] 土屋誠司, 佐竹純二, 近間正樹, 上田博唯, 大倉計美, 蚊野浩, 安田昌司. Tv 番組推薦システムの構築とその有用性の検証. 情報処理学会研究報告（ヒューマンインタフェース研究会）, Vol. 2006, No. 3, pp. 95–102, 2006.

[12] Tomohiro Tsunoda and Masaaki Hoshino. Automatic metadata expansion and indirect collaborative filtering for tv program recommendation system. *Multimedia Tools and Applications*, Vol. 36, No. 1-2, pp. 37–54, 2008.