

潜在トピックを利用した協調フィルタリングにおける トピック情報源の違いに関する分析

西村 章宏¹ 土方 嘉徳^{1,a)}

概要：大量の情報の中から、ユーザが必要とする情報を自動で選別するための一手法として協調フィルタリング (CF) が存在する。この CF の中でも Matrix Factorization (MF) は、評価値の欠損が多い現実のデータセットに対して優れた結果を出している。ただし、MF は評価値の欠損に強いものの、評価値が極端に少ない場合には適用することが困難である。この問題に対して、評価値だけでなくテキスト情報も利用するアプローチが近年注目されている。本研究では、このアプローチの先駆けとなったモデルである Collaborative Topic Regression (CTR) に注目する。CTR では、トピックを抽出する情報源として、アイテムに関するテキスト情報を用いている。本研究では既存の CTR モデル (iCTR と呼ぶ) をベースとして、ユーザに関するテキスト情報を利用する uCTR モデルを提案する。この両者とベースラインである MF に対して推薦結果の正確性 (再現率) と利便性 (被覆率・多様性) に関して定量的な比較を行う。この際、極端に評価値が欠損している場合でも CTR が MF より有効にはたらくことを確認するため、評価値行列のスパース性の程度 (以下スパース率) を変化させて実験を行う。

1. はじめに

ユーザの情報獲得を支援する方法の一つに推薦システムがある。推薦システムに関する研究は広く行われており、それを実現する方法は、ユーザの興味・嗜好に関する情報とアイテムの内容に関する情報を元に推薦を行う内容ベースフィルタリングと、対象ユーザと嗜好の似ている他者が好むアイテムを推薦する協調フィルタリング (CF) に大きく分けることができる [1]。このうち CF は、ユーザやアイテムの内容に関する特徴量を作成する必要がないという簡便さから、多くの商用推薦システムで用いられている。

CF ではユーザがアイテムを好む程度 (評価値) が行列形式で表現される。この行列は評価値行列と呼ばれる。ただし、実システムやサービスで得られる評価値行列には欠損値が多く存在しており、評価値不足により推薦が上手く行えないことがある (スパース性の問題)。そのため、より少ない評価値から効率的に未知の評価値を推定する手法として、Matrix Factorization (MF) [2] が存在する。

MF はユーザとアイテムをより低次元な潜在特徴空間へマッピングする手法であり、このユーザとアイテムの潜在因子ベクトルを用いて元の評価値を近似することができる。このため、評価値行列中に欠損値が含まれていたとしても、

より欠損値の少ない評価値行列を推定することができ、実データに対しても優れた推薦を行うことができる [3]。ただし、MF による欠損値への対処にも限度がある。評価値が極端に少ない新規アイテムや新規ユーザに対してはうまく適用することができない。これは一般にコールドスタート問題と呼ばれる [1], [4]。このような評価値が極端に不足する問題は、知名度が低いニッチなアイテムが多く存在するドメインにおいてもよく見られる。

この問題に対して、近年の MF の研究では、評価値行列だけでなくアイテムに関するテキスト情報も利用している [5], [6], [7], [8]。これらの研究と MF との大きな違いは、アイテムに関するテキスト情報から潜在トピック (以降は単にトピックと呼ぶ) を抽出し、ユーザとアイテムを表現する潜在因子ベクトルにこのトピックを反映させる点である。抽出したトピックに誘起して潜在因子ベクトルが生成されるため、評価値が全く存在しない新規アイテムに対しても、評価値が存在する他のアイテムとのトピックの類似関係を元に評価値を推定することができる [5]。

本研究では、このテキストから抽出したトピックを利用するアプローチの先駆けとなったモデルである Collaborative Topic Regression (CTR) [5] に注目する。CTR では、トピックを抽出する情報源として、アイテムに関するテキスト情報を用いている。一方で、ユーザに関するテキスト情報を用いることに関しては深く言及されていない。ユーザ

¹ 大阪大学大学院 基礎工学研究科

^{a)} hijikata@sys.es.osaka-u.ac.jp

に関するテキスト情報から抽出されるトピックは、主にユーザの種類（ユーザ層）であると考えられる [9]。そこで我々は、このユーザ層をユーザとアイテムを表現する潜在因子ベクトルに反映させることを考える。

すなわち、本研究では既存の CTR モデル（iCTR と呼ぶ）をベースとした uCTR モデルを提案し、この両者とベースラインである MF に対して、推薦結果の正確性（再現率）と利便性（被覆率・多様性）の両方の観点から、定量的な比較を行う。この際、極端に評価値が欠損している場合でも CTR が MF より有効にはたらくことを確認するため、評価値行列のスパース性の程度（以下スパース率）を変化させて実験を行う。

2. 関連研究

協調フィルタリング（CF）を用いた推薦手法は大きく 2 種類に分けられる [1], [10]。1 つは、評価値から直接ユーザ間またはアイテム間の比較を行うメモリベース法である。ピアソン相関でユーザ間またはアイテムの類似度を測る手法がよく知られている [11], [12]。もう 1 つは、評価値を利用して推薦を行うためのモデルを学習するモデルベース法である。評価値を予測する関数のパラメータを回帰分析や因子分析等のアプローチで学習するモデル [2], [13] や、ユーザ・アイテム・評価値をそれぞれ確率変数と捉えて同時分布を学習するモデル [14] などが存在する。中でも、元の評価値行列を複数の低次元の行列に分解し、ユーザやアイテムが持つ潜在特徴を利用するモデルが高い性能を発揮することが知られている [3], [15]。これらは、行列因子分解モデルと呼ばれる。行列因子分解モデルの中でも、 $R \approx U^T V$ となるように元の評価値行列 R を近似する行列 U, V を推定するモデル [2] は、実装が容易かつ高性能であることから広く利用されており、一般的に Matrix Factorization (MF) と呼ばれている。

このような潜在特徴に関しては、自然言語処理の分野でも活発に研究が行われており、近年ではトピックモデルと呼ばれる手法が注目されている。これは、文書の生成過程を潜在変数を交えて階層的に表現し、各階層間の関係を確率的にモデル化したものである。これにより文書集合から容易にトピック（文書中の潜在因子）を抽出することが可能となった。トピックモデルの中で最も基本的なモデルは、Blei らが提案した Latent Dirichlet Allocation (LDA) [16] であり、モデルの階層構造が比較的シンプルでありながら強力なトピック抽出機能を有している。

近年では、CF に評価値以外の情報を利用することで、コールドスタート問題やスパース性の問題に対処しようとする研究が活発に行われている。これらの研究は CF と内容ベースフィルタリングを組み合わせたハイブリッド法とみなせる。ハイブリッド法には、両方のモデルの結果を混合して提示する方法 [17] や、ユーザプロファイルやアイテ

ム内容をベクトルや確率分布の形式に変換して利用する方法 [18], [19] 等が存在する [1]。その中でも、両方のモデルに前後関係が無く、同等のレベルでモデル化を行う完全結合 [1] に注目する。この完全結合のモデルは階層的な確率モデルとして表現することが可能であるため [20], [21], [22]、先述のトピックモデルを自然な形で取り入れることができる。そのため、CF とトピックモデルを組み合わせたアプローチが研究され始めている [5], [6], [7], [8]。

我々は CF とトピックモデルを融合させたモデルの先駆けとなった collaborative topic regression (CTR) [5] に注目する。CTR は Wang と LDA の提案者である Blei が提案したものである。シンプルなモデルでありながらも、優れた推薦性能を実現している。本研究では、CTR を CF とトピックモデルを融合させたモデルのベースラインとして捉え、トピックを抽出する情報源の違いに着目して推薦結果に与える影響の分析を行う。具体的には、CTR はアイテムに関するテキスト情報からアイテムの潜在トピックを抽出して潜在因子ベクトルに反映させていたが、本研究ではユーザに関するテキスト情報からアイテムの潜在トピックを抽出して潜在因子ベクトルに反映させる。また、従来研究ではスパースさに対するロバスト性については分析されてなかったが、本研究ではスパース率を変化させて、推薦性能に関する実験を行う。また、推薦性能を正確性だけでなく利便性の観点からも評価する。

本研究の貢献をまとめると以下のとおりである。

- (1) CF とトピックモデルを融合させたモデルにおいて、ユーザに関するテキスト情報を用いたモデルを提案し、情報源の違いが推薦結果に与える影響について分析した。
- (2) 性能評価には正確性に関する評価指標に加えて、推薦の利便性を表す被覆率と多様性に関する評価指標も用いた。
- (3) 上記モデルの性能評価において、スパースさに対するロバスト性について分析した。

3. 推薦モデル

CTR モデルにおいて、トピックを抽出する情報源の違いに着目する。具体的には、アイテムに関するテキスト情報からトピックを抽出するモデルと、ユーザに関するテキスト情報からトピックを抽出するモデルが考えられる。この章では、これらのモデルの概要を説明する。

3.1 Collaborative Topic Regression (item-oriented)

CTR モデル [5] は、協調フィルタリングにトピックモデルを融合させるアプローチの先駆けとなったモデルである。図 1 に示したグラフィカルモデルにおいて、上半分（左端の α から右端の β までの横の並び）はトピックモデルで

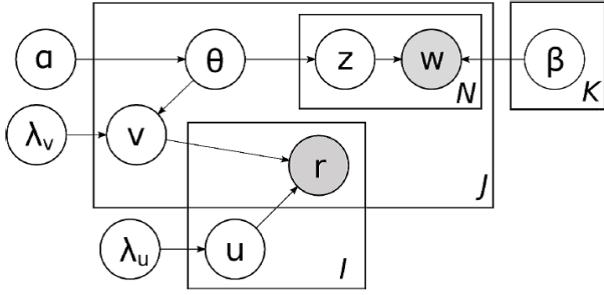


図 1 Collaborative Topic Regression (item-oriented) のグラフィカルモデル

ある LDA[16] の生成過程とほぼ同じであり、下半分 (v や λ_v 以下の部分) は MF に相当する処理を表している。この MF の部分は、確率的モデルである Probabilistic Matrix Factorization (PMF) [23], [24] がベースとなっている。図中の各変数については以下に続く節で詳しく説明するが、ポイントのみ押さえて説明すると、CTR は LDA 部分の θ と PMF 部分の v との間に依存関係を与えたモデルだということが分かる。この依存関係により、アイテムに関する文書から抽出したアイテムを特徴づけるトピック比率 θ をアイテムの潜在因子 v へ反映させている。

3.1.1 モデル説明

最初に、このモデルに登場する変数に関して説明を行う。添え字として、各ユーザを i (総数 I)、各アイテムを j (総数 J)、各トピックを k (総数 K) と表す。実際にデータとして得られる観測変数は、アイテムに関するテキスト中の単語 (トークン) 列 w_j 、ユーザのアイテムに対する評価値 r_{ij} である。今回行うタスクにおいて、評価値は $r_{ij} \in \{0, 1\}$ の二値である。 $r_{ij} = 0$ は欠損値であり、ユーザ i がアイテム j に興味を持っていない、もしくは存在を知らないことを意味する。また、全アイテムの w_j を集め、その中でユニークな単語の集合 (語彙集合) を定義し、各語彙を表す添字を y (総数 Y) とする。加えて、 w_j の各トークンを指す添え字を n (総数 N) とする。潜在変数である z_{jn} は、トークン w_{jn} に対して割り当てられるトピックである。

これら観測変数と潜在変数の他に、モデルを構成する要素としてパラメータが存在する。パラメータは確率変数の生成確率を表現する変数であり、これを推定することがモデルの学習であるといえる。トピック比率 θ_j は $\sum_k \theta_{jk} = 1$ を満たし、各アイテムが持つトピックの傾向を表している。この θ_j の比率に基づき、多項分布から確率的に z_{jn} が生成される。語彙比率 β_k は $\sum_y \beta_{ky} = 1$ を満たし、トピック k における語彙の出現確率を表している。 u_i はユーザの特徴を表現する K 次元の潜在因子ベクトル、 v_j はアイテムの特徴を表現する K 次元の潜在因子ベクトルである。

パラメータの分布を制御する変数として、ハイパーパラメータが存在する。 λ_u, λ_v は、それぞれ u_i, v_j を生成する正規分布の分散を制御するハイパーパラメータであり、各

u_i, v_j に対して平滑化 (要素値が極端な値を取ることを抑制) を行う効果がある。 α は、 θ_j を生成するディリクレ分布のハイパーパラメータであり、アイテム全体における生成されるトピックの偏り具合を制御する。なお元論文 [5] に従い、これらのハイパーパラメータはいずれもスカラーである。

MF を確率モデルとして一般化した PMF[23], [24] では、評価値は (式 1) により生成される。

$$\begin{aligned} u_i &\sim N(0, \lambda_u^{-1} I_K) \\ v_j &\sim N(0, \lambda_v^{-1} I_K) \\ r_{ij} &\sim N(u_i^T v_j, c_{ij}^{-1}) \end{aligned} \quad (1)$$

ただし、 I_K は K 次元の単位行列、 $N(\mu, \sigma)$ は正規分布を表す。式中の c_{ij} は、既知の評価値 r_{ij} の影響を制御するハイパーパラメータである。欠損値となる要素に対しては、該当する要素を $c_{ij} = 0$ と設定することでその影響を無視する。特に $c_{ij} = 1$ の場合、PMF の最大事後確率 (MAP) 推定は、MF の目的関数を最小化 (式 2) した結果と一致する。

$$\min_{U, V} \sum_{ij} (r_{ij} - u_i^T v_j)^2 + \lambda_u \|u_i\|^2 + \lambda_v \|v_j\|^2 \quad (2)$$

CTR では、 c_{ij} は次の様に定義する。

$$c_{ij} = \begin{cases} a & (r_{ij} = 1 \text{ のとき}) \\ b & (r_{ij} = 0 \text{ のとき}) \end{cases} \quad (3)$$

次に、CTR の生成過程を説明する。

1. 各ユーザ i において、正規分布 $N(0, \lambda_u^{-1} I_K)$ から u_i を生成。
2. 各アイテム j において、
 - a. ディリクレ分布 $Dir(\alpha)$ から θ_j を生成。
 - b. 正規分布 $N(0, \lambda_v^{-1} I_K)$ から K 次元ベクトル ϵ_j を生成し、 $v_j = \epsilon_j + \theta_j$ とする。
 - c. 各トークン w_{jn} において、
 - i. 多項分布 $Mult(\theta)$ から、トピック z_{jn} を生成。
 - ii. 多項分布 $Mult(\beta_{z_{jn}})$ から、トークン w_{jn} を生成。
3. 各ユーザ・アイテムのペア (i, j) において、正規分布 $N(u_i^T v_j, c_{ij}^{-1})$ から r_{ij} を生成。

3.1.2 モデルの学習

この CTR モデルのパラメータ (u, v, θ, β) の学習は、元論文 [5] に従いモデルの事後確率 L (式 4) を最大化するように反復的に最適化を行う。各パラメータの更新は、更新するパラメータ以外を固定して事後確率 L を最大化し、更新したいパラメータを最適化する coordinate ascent 法 [27] により行う。

$$L = - \sum_i \sum_j c_{ij} (r_{ij} - u_i^T v_j)^2 + \sum_j \sum_n \log(\sum_k \theta_{jk} \beta_{k,w_{jn}}) \text{ リストとしてユーザへ提示する.} \\ - \lambda_u \sum_i \|u_i\|^2 - \lambda_v \sum_j \|v_j - \theta_j\|^2 \quad (4)$$

u, v の更新は、(式 5) により行う。式中の U, V はそれぞれ列ベクトル u_i, v_j を並べた行列、 C^i, C^j は対角行列でそれぞれ $C_{dd}^i = c_{id}, C_{dd}^j = c_{dj}$ 、 r_i, r_j はベクトルでそれぞれ $r_i = (r_{i1}, r_{i2}, \dots, r_{iJ}), r_j = (r_{1j}, r_{2j}, \dots, r_{Ij})$ である。各 u_i に対して事後確率 L を大きくする勾配方向へ u_i を微小量ずつ変化させていく。各 v_j に対しても同様に変化させていく。

$$u_i = (VC^i V^T + \lambda_u I_K)^{-1} VC^i r_i \quad (5) \\ v_j = (UC^j U^T + \lambda_v I_K)^{-1} (UC^j r_j + \lambda_v \theta_j)$$

θ の更新において、事後確率 L (式 4) の第 2 項目を直接求めることが非常に困難である。そこで、(式 4) から θ_j を含む項を取り出し、アイテム毎に分離して $L(\theta_j)$ とする。次に、 $\phi_{jnk} = q(z_{jn} = k)$ を定義し、Jensen の不等式を利用して (式 6) の様に変形する。

$$L(\theta_j) = -\lambda_v \|v_j - \theta_j\|^2 + \sum_n \log(\sum_k \frac{\theta_{jk} \beta_{k,w_{jn}} \phi_{jnk}}{\phi_{jnk}}) \\ \geq -\lambda_v \|v_j - \theta_j\|^2 \quad (6) \\ + \sum_n \sum_k \phi_{jnk} (\log(\theta_{jk} \beta_{k,w_{jn}}) - \log(\phi_{jnk})) \\ = L(\theta_j, \phi_j)$$

この $L(\theta_j, \phi_j)$ は、本来求めたい $L(\theta_j)$ の下限となっている。また、最適な ϕ_{jnk} は $\phi_{jnk} \propto \theta_{jk} \beta_{k,w_{jn}}$ を満たす。 θ_j の最適化に関しては、解析的に行うことができないため、 θ_j が simplex であることを制約条件に projection gradient 法 [27] を利用して探索的に最適化を行う。

最後に β の更新は、(式 7) により行う。

$$\beta_{kw} \propto \sum_j \sum_n \phi_{jnk} \delta(j, n, w) \quad (7) \\ \delta(i, n, w) = \begin{cases} 1 & (w_{jn} = w \text{ のとき}) \\ 0 & (w_{jn} \neq w \text{ のとき}) \end{cases}$$

以上のパラメータ (u, v, θ, β) の更新処理を、事後確率の変化が微小になり収束したと見なせるまで繰り返し実行する。

3.1.3 評価値の予測と推薦

モデルの学習が完了した後、推定したパラメータ $\hat{u}, \hat{v}$ を用いて (式 8) から予測評価値 $\hat{r}$ を計算することができる。この予測評価値は実数値となるため、値の大きい順にランキングを作成することができる。推薦時には、あるユーザ i に関連する評価値 r において、値が 0 の要素 (i, j) を抽出し、それらの予測評価値 $\hat{r}$ を求める。そして、予測評価値 $\hat{r}_{ij}$ の値が大きい順にアイテム j を列挙したものを、推薦

$$r_{ij} \approx \hat{r}_{ij} = u_i^T v_j \quad (8)$$

3.2 Collaborative Topic Regression (user-oriented)

前節の CTR モデルでは、トピックをアイテムに関する情報源から抽出していたが、この情報源をユーザに関する情報源に変更した場合について考える。すなわち、図 1 に示したグラフィカルモデルにおいて、 u と v および関連する $\lambda_u, \lambda_v, I, J$ を入れ替えたモデルを構築する (以降、このモデルを uCTR、前節のオリジナルのモデルを iCTR と呼ぶ)。uCTR と iCTR の違いは、トピックを抽出する情報源が異なる点である。このため、uCTR モデルでは抽出される K 次元のトピック分布 θ の内容もユーザに関するトピックとなり、このトピックの影響を受ける潜在因子ベクトル u, v はユーザに関する潜在因子となる。uCTR のパラメータ学習に関しては、3.1.2 と同様に最大事後確率 (MAP) 推定により行う。

4. 評価方法

推薦結果の正確性と利便性を評価する。交差検定により評価指標の値を算出して定量的に評価を行う。

4.1 評価方法の概要

推薦システムの性能を調べるため、推薦の正確性を表す指標として再現率、推薦の利便性を表す指標として被覆率と多様性の評価を行う。本評価では不特定多数のユーザによる評価データを利用するため、交差検定によるオフライン評価 [4] を行う。具体的には、データを 5 分割して訓練データ 4 つとテストデータ 1 つに分け、テストデータを擬似的にユーザが好む未知のアイテム (正解データ) とみなす。訓練データを利用してモデルの学習を行った後、上位 N 個の推薦アイテムを推薦リストとして獲得する。推薦リストと正解データを比較して評価指標の値を算出する。

4.1.1 評価指標

正確性と利便性について、本研究で利用する評価指標を紹介する。正確性に関する評価としては、適合率 (精度) と再現率が代表的であるが、元論文 [5] や類似の研究 [7], [23] に倣い、再現率を利用する。インターネット上には非常に多くのコンテンツ (アイテム) が存在しているが、これらのアイテムをユーザが網羅的に閲覧することは不可能である。そのためオフライン評価では、ユーザがまだ閲覧していないが閲覧すれば好むアイテムが好きでないと扱われる。この問題による影響を抑えるため、正確性の評価として適合率ではなく再現率を採用することがある。また、利便性の評価指標としては、推薦システムが偏りなく様々なアイテムを実際に推薦できているかを表す被覆率と、ユー

ザ毎に個別化された推薦ができていないかを表す多様性を採用する。以下に詳細を示す。

再現率

再現率 (Recall) は以下の式で表される。

$$Recall = \frac{|\{\text{推薦結果の要素}\} \cap \{\text{正解の要素}\}|}{|\{\text{正解の要素}\}|} \quad (9)$$

最終的なシステムの再現率は、全ユーザの Recall の平均を取ることで求める。

被覆率

被覆率は、全アイテムの中で実際に推薦可能なアイテム (テストセット中で推薦されたアイテム) がどの程度存在するかを表したものである。本研究では、被覆率の指標の 1 つである Catalogue Coverage (CC) [25] を用いる。あるユーザ i に提示される推薦リストを L_i とすると、CC は次式で求められる。

$$CC = \frac{|\cup_{i=1 \dots I} E(L_i)|}{|B|} \quad (10)$$

$E(L_i)$ はあるリスト L_i 内のアイテム集合への写像、 B はデータセット中のアイテム集合である。CC が低いほど、全アイテム中の一部のアイテムしか推薦システムは扱えていないと判断できる。

多様性

多様性の指標には、二人のユーザに注目した時に、そのユーザの推薦リストに含まれるアイテムがどれほど違っているかを測定する Inter-User Diversity (IUD) [26] を用いる。IUD は次式で求められる。

$$IUD = \frac{1}{|U|C_2} \sum_{u1}^U \sum_{u2}^U d_{u1,u2} \quad (11)$$

$$d_{u1,u2} = 1 - \frac{|E(L_{u1}) \cap E(L_{u2})|}{S}$$

ただし、 $S = |E(L_{u1})| = |E(L_{u2})|$ である。IUD は $[0, 1]$ に正規化されたユーザ間の距離の平均と捉えることができ、0 に近いほどユーザ間の推薦結果に差は無く、1 に近いほど差が存在することを意味している。

5. 実験

iCTR, uCTR, MF の比較実験を行う。使用するデータセットは、元論文 [5] に倣い、ソーシャルブックマークサービス CiteULike*1 におけるユーザがブックマークを行った論文・科学記事である。我々は、CiteULike において 2014 年 9 月から 11 月の間にブックマークを行ったユーザとそのユーザが登録したすべての論文・科学記事を、CiteULike を直接クロールすることで収集した。また、ユーザに関してはそのプロフィール情報を、論文・科学記事に関してはそのタイトルと要約情報も収集した。

*1 Web 上に存在する論文や Web ページのブックマークを行い、それらを整理・共有できるサービス

5.1 前処理

CiteULike で扱っている論文や科学記事の大半は英語で書かれている。そのため、実験でも英語のテキストを対象とする。英語の文章に対しては、正規化 (小文字に統一、ステミング) を行った後、ストップワード (he, the, this など) を除去することが一般的である。本実験では、ステミングに関しては TreeTagger*2 を利用し、ストップワードの除去に関しては SlothLib*3 を利用した。

5.2 実験設定

本実験では、CiteULike のデータを用いる。ユーザの評価値は記事ブックマーク (ある論文・記事をブックマークしたかどうか) である。そのため、ユーザがあるアイテム (論文・記事) に関心がある場合には 1、関心が無いもしくは存在を知らない場合には 0 となる。iCTR では、トピック抽出元のテキストにはアイテムの “article title” と “abstract” を用いる。uCTR では、トピック抽出元のテキストにはユーザの “profile” と “interest” (ユーザが興味のある分野をキーワードで列挙したもの) を用いる。

また、用いたデータセットの評価値行列のスパース率を計算したところ、99.926 % と非常にスパースであった。我々はスパース率による結果の変化を調べるため、被ブックマーク数がある閾値以下のアイテムを除外することでスパース率を低下させる処理を行った。本実験では、被ブックマーク数が 2, 4, 6, 8, 10 未満のアイテムを除外した際のそれぞれの結果 (順に nlt=2, 4, 6, 8, 10) を求め、スパース率の増減によるモデル毎の性能の変化についても分析する。なお、nlt=2, 4, 6, 8, 10 におけるスパース率は順に、99.926, 99.894, 99.827, 99.754, 99.678 (%) であった。スパース率の変化で見ると非常に小さく見えるが、評価値が存在する率として見ると、nlt=2 で 0.0743%, nlt=10 で 0.3241% になっており、約 4.36 倍の差がある。

モデルのハイパーパラメータ設定は、グリッドサーチにより行った。潜在因子ベクトルの次元に相当するトピック数 K については、特徴の表現力がある程度維持しつつ、特徴量の次元数が可能な限り小さい値となるように、段階的に設定して評価指標の値を確認した。具体的には、トピック数 ($K = 5, 10, 30, 50, 100, 200$) において評価指標の値の向上が微小になった $K = 50$ を用いた。潜在因子ベクトルの平滑化度合いを調整する λ_v, λ_u については、元論文 [5] を参考に iCTR では $\lambda_v = 1, 10, 100, \lambda_u = 0.1, 1, 10$, uCTR では λ_v と λ_u の値を入れ替えた設定で、それぞれの組み合わせを調べた際の最も良い結果 (iCTR では ($\lambda_v = 100, \lambda_u = 0.1$), uCTR では ($\lambda_v = 0.1, \lambda_u = 100$)) を用いた。残りのハイ

*2 <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

*3 <http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/English.txt>

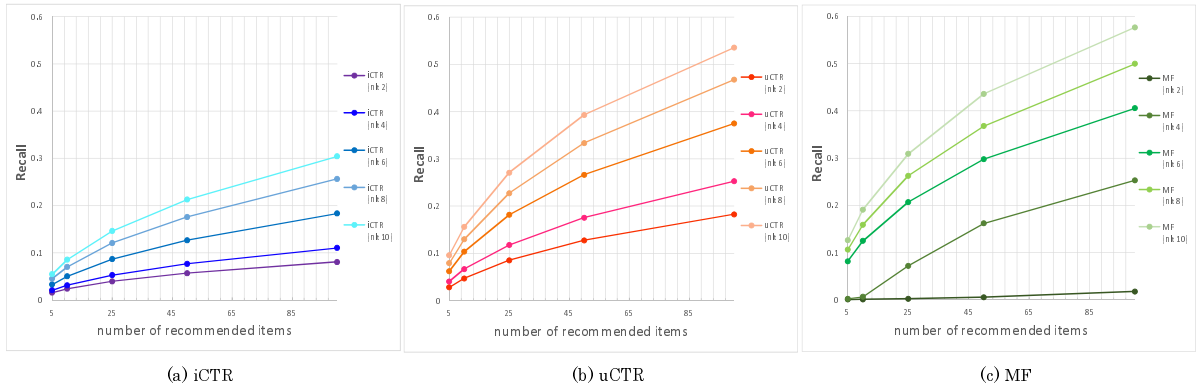


図 2 Recall, モデル・スパース率毎の比較 (nlt=2 が最もスパース)

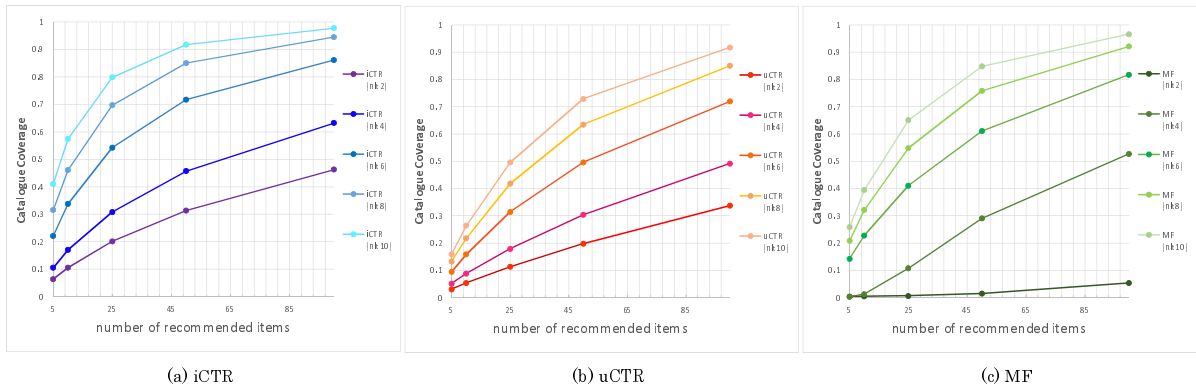


図 3 Catalogue Coverage, モデル・スパース率毎の比較 (nlt=2 が最もスパース)

パラメータに関しては、元論文 [5] と同じ値を用いた ($a = 1, b = 0.01, \alpha = 1$)。パラメータ θ と β の初期値に関しては、事前にデータセットに用いる文書集合を使って LDA モデルの学習を行い、その結果得られる θ と β の値を利用する (LDA の学習は Collapsed Gibbs Sampling[28] で 500 回の反復を行った)。その他のパラメータに関しては、乱数により初期化を行う。なお、MF の実装は、元論文 [5] の方法を参考に、CTR モデルのハイパーパラメータを ($\lambda_v = 0.1, \lambda_u = 0.1, a = 1, b = 0.01, \alpha = 1$) と設定し、かつ θ の値を全て 0 に固定したものを利用する。

5.3 結果・考察

5.3.1 再現率

正確性の指標である再現率 (Recall) の結果を図 2 に示す。このグラフは iCTR, uCTR, MF のモデル間の比較、および 5 段階に分けたスパース率間での比較を行っている。横軸は推薦リストに含めるアイテムの数 (推薦数) である。まず同一モデルにおけるスパース率の変化による影響に注目すると、どのモデルにおいてもスパース率が高いほど (nlt の値が小さいほど)、Recall の値は小さくなる傾向が見られる。

ここで興味深いことに、両 CTR と MF とではスパース率の変化に伴う Recall の変化量が大きく異なっている。推薦数が 100 となる値 (グラフ右端) に注目すると、MF で

はスパース率が最大の時 (nlt2) と最小の時 (nlt10) の間で Recall の差は約 0.55 となっている。対して、iCTR はスパース率が最大の時と最小の時の間で Recall は約 0.2 しか変わらず、uCTR は約 0.35 しか変わらない。MF に比べるとスパース率による影響が小さく抑えられている。これらのことから、CTR は MF に比べてデータの欠損に対してロバストであることが分かる。

一方で、データのスパース率が小さい場合は、MF が最も性能が良くなっている。これは、CTR でテキストから抽出されたトピックの情報よりも、評価値行列に存在する潜在的な情報の方がよりリッチであり、トピックの情報がノイズとなってしまっているためと考えられる。また、iCTR と uCTR との間にも明らかな差が存在している。特に、iCTR はスパース率が低い場合には、uCTR よりも顕著に Recall が低くなっている。この原因を解明することは非常に困難であるが、アイテムのトピックを利用するよりも、ユーザのトピックを利用する方が、推薦の正確性を向上させる可能性があることを示す結果と言える。

以上の結論としては、CTR はどちらも評価値からの情報不足をトピックにより補うことができている、スパース性の影響を抑えることができるといえる。特に uCTR においては、より推薦の正確性を向上させる可能性があることを確認できた。ただし、評価値情報が十分である場合には、かえってノイズとなる可能性があることに注意する必

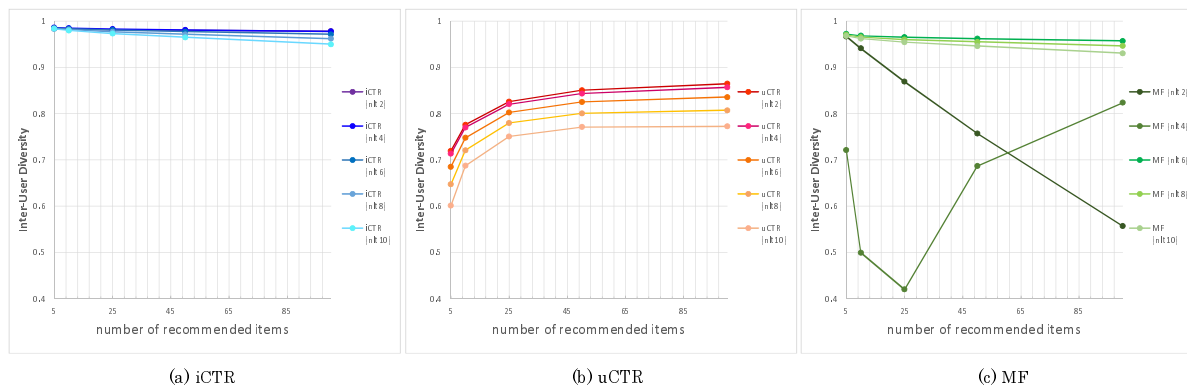


図 4 Inter-User Diversity, モデル・スパース率毎の比較 (nlt=2 が最もスパース)

要がある。

5.3.2 被覆率

利便性の指標に関しては、最初に被覆率を表す Catalogue Coverage (CC) の結果を図 3 に示す。同一モデルにおけるスパース率の変化による影響は、Recall の傾向と似たものとなった。すなわち、スパース率が高いほど (nlt の値が小さいほど)、CC の値は小さくなる傾向が見られた。スパース率が高い場合と低い場合の CC の値の差は、推薦数が増えるにつれ、大きくなった。

手法間で比較を行うと、iCTR がいずれのスパース率においても、またいずれの推薦数においても、他の 2 手法より優れた結果となっている。MF の結果は uCTR の結果より優れているが、スパース率が非常に低い場合 (nlt2) では、極端に CC が低くなっており uCTR の CC よりも低くなっている。このことより、トピックの導入はスパース率が高い場合に、CC が極端に減少する現象を抑制する効果があることが分かる。推薦を行うための元となるユーザやアイテムに関する情報 (評価値内の潜在的な情報も含む) が多いほど、モデルのパラメータはより多様なユーザとアイテムを表現できるように学習され、結果として多様なパターンの推薦を行えるからだと考えられる。

5.3.3 多様性

また別の利便性の指標として、多様性を表す Inter-User Diversity (IUD) の結果を図 4 に示す。このグラフを見ると、iCTR では、スパース率を変化させても、またいずれの推薦数でも、IUD は非常に高い値で安定しており、3 つの手法の中で最も良い結果となった。したがって、ユーザごとに個別化された多様な推薦が実現できていると言える。uCTR は、iCTR や MF (スパース率が低い場合) より悪い結果となった。ただし、それでも 0.8 前後の値を示しており、特に推薦数が多くなれば、ある程度高い個別化が行われていることが分かる。IUD の変化に注目すると、それほど大きな差ではないが、スパース率が高いほど IUD も高くなっている。推薦数が 5 から 50 の間に IUD が上昇している傾向が見られるが、これは推薦数が少ない時にはメ

ジャーなアイテムを推薦し、推薦数が増えると個別化されたアイテムを推薦しているからだと考えられる。

MF は、スパース率が低い場合は、iCTR に匹敵するほどの良い結果となった。しかし、一部の結果 (MF (nlt2) と MF (nlt4)) が他とは異なる挙動を示していることが見て取れる。MF (nlt2) は、急激に単調減少している様子が見られるが、これに関しては CC が非常に小さいことから、どのユーザに対しても同じアイテムしか推薦できていないためだと考えられる。MF (nlt4) は、複雑な挙動を示しており、推薦数 25 以降では IUD が上昇する様子が見られる。我々はこの原因を特定することはできなかったが、CC の値がこの複雑さを生み出している可能性があると思われる。以上より、MF ではスパース率が高くなると IUD の値が急激に減少したり不安定になる傾向があるが、トピックを導入することでその現象を起こさないようにすることができると言える。

5.4 結果のまとめ

一般に推薦の正確さを重視するのであれば、テキストから抽出したトピックを利用するのではなく、オリジナルの MF を用いるほうが良い。しかし、データセットのスパース率が高い場合は、uCTR を用いるほうが良い。MF は、スパース率によって大きくその正確性が変わってしまうため、システム運用後のスパース率の値が想定できない場合は、uCTR を用いるほうが無難と言える。また、多様なアイテムが推薦されるようにするためには iCTR を用いるほうが良い。特に推薦数が少ない場合には有効である。また、スパース率が高い場合は、MF の使用は避けた方が良い。MF はスパース率が高くなると、推薦するアイテムがある特定のものに偏ってしまうからである。最後に、ユーザごとに個別化された推薦ができることを重視するのであれば、iCTR を用いるほうが良い。iCTR は、推薦数やスパース率が変わっても、安定して高い IUD を保つからである。

6. おわりに

本研究では、協調フィルタリングとトピックモデルを組み合わせた代表的モデルである Collaborative Topic Regression (CTR) モデルに注目し、アイテムに関するテキストだけでなくユーザに関するテキストを利用することを考え、これらとベースラインである Matrix Factorization (MF) との比較を行う実験を行った。実験では、評価値の欠損にどれだけ対応できるかに重点を置き、評価値行列のスパース率を変化させた際の iCTR, uCTR と MF の性能を調べた。それぞれの手法を総合的な観点から評価を行うために、正確性に関する評価指標として再現率を、利便性に関する評価指標として被覆率と多様性を用いた。

実験の結果、どちらの CTR モデルも MF に比べて評価値の欠損による影響が小さくロバストであることが示せた。iCTR と uCTR のどちらが優れているかに関しては一概に決めることは難しく、iCTR では個別化の高い推薦、uCTR では正確性の高い推薦を行う傾向が見られた。また、MF はスパース率が低い場合は最も正確性の高い推薦が行えたが、スパース率に対して極めて敏感であるため、扱いには注意が必要であることが分かった。

今後の課題としては、ハイパーパラメータ間の交互作用、すなわちトピック数 K とモデルの複雑さを調整する λ_u, λ_v の設定を変化させた場合に、非常にスパースなデータセットとそうでないデータセット間に違いが生じるか調査する予定である。

謝辞

本研究は科研費 (15K12150) の助成を受けたものである。

参考文献

- [1] 神島 敏弘: 推薦システムのアルゴリズム (3), 人工知能学会誌 Vol.23, No.2, pp.248-263 (2008)
- [2] Koren, Y., Bell, R. and Volinsky, C.: Matrix factorization techniques for recommender systems, Computer 42(8), pp.30-37 (2009)
- [3] Paterek, A.: Improving Regularized Singular Value Decomposition for Collaborative Filtering, Proc. ACM SIGKDD Cup and mWorkshop (2007)
- [4] 土方 嘉徳: 推薦システムのオフライン評価手法, 人工知能学会誌 Vol.29, No.6, pp.658-689 (2014)
- [5] Wang, C. and Blei, D.M.: Collaborative topic modeling for recommending scientific articles, Proc. ACM SIGKDD, pp.448-456 (2011)
- [6] Ding, X., et al.: Celebrity Recommendation with Collaborative Social Topic Regression, Proc. IJCAI (the 23th international joint conference on Artificial Intelligence), pp.2612-2618 (2013)
- [7] Purushotham, S., Liu, Y. and Kuo, C.-C.J.: Collaborative Topic Regression with Social Matrix Factorization for Recommendation Systems, Proc. ICML, pp.759-766 (2012)
- [8] Kim, Y. and Shim, K.: TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation, Journal of Information Systems, Vol.42, pp.59-77 (2014)
- [9] 西村 章宏, 土方 嘉徳, 三輪 祥太郎, 西田 正吾: 一般ユーザの観点に基づく Twitter からの人物関係の可視化と事例の考察, 情報処理学会論文誌, Vol.56, No.3, pp.972-982 (2015)
- [10] Su, X. and Khoshgoftaar, T.M.: A survey of collaborative filtering techniques, Journal of Advances in Artificial Intelligence, Vol.2009, No.4 (2009)
- [11] Resnick, P., et al.: GroupLens: An open architecture for collaborative filtering of Netnews, Proc. ACM CSCW'94, pp.175-186 (1994)
- [12] Sarwar, B., et al.: Item-based Collaborative Filtering Recommendation Algorithms, Proc. ACM WWW, pp.285-295 (2001)
- [13] Canny, J.: Collaborative filtering with privacy via factor analysis, Proc. ACM SIGIR, pp.238-245 (2002)
- [14] Hofmann, T. and Puzicha, J.: Latent class models for collaborative filtering, Proc. IJCAI'99, pp.688-693 (1999)
- [15] Cremonesi, P., Koren, Y. and Turrin, R.: Performance of Recommender Algorithms on Top-N Recommendation Tasks, Proc. ACM RecSys, pp.39-46 (2010)
- [16] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, The Journal of Machine Learning Research, Vol.3, pp.993-1022 (2003)
- [17] Basu, C., et al.: Technical paper recommendation: A study in combining multiple information sources, Journal of Artificial Intelligence Research, Vol.14, pp.231-252 (2001)
- [18] Melville, P., Mooney, R. J. and Nagarajan, R.: Content-booster collaborative filtering for improved recommendations, Proc. AAAI, pp.187-192 (2002)
- [19] Yu, K., et al.: Collaborative ensemble learning: Combining collaborative and content-based information filtering via hierarchical bayes, Proc. UAI'03, pp.616-623 (2003)
- [20] Schein, A. I., et al.: Methods and Metrics for Cold-Start Recommendations, Proc. ACM SIGIR'02, pp.253-260 (2002)
- [21] Popescul, A., et al.: Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments, Proc. UAI'01, pp.437-444 (2001)
- [22] B. M. Kim and Q. Li.: Probabilistic model estimation for collaborative filtering based on items attributes, Proc. IEEE/WIC/ACM WI'04, pp.185-191 (2004)
- [23] Hu, Y., Koren, Y. and Volinsky, C.: Collaborative Filtering for Implicit Feedback Datasets, Proc. IEEE ICDM, pp.263-272 (2008)
- [24] Salakhutdinov, R. and Mnih, A.: Probabilistic Matrix Factorization, Proc. NIPS 21 (2008)
- [25] Ge, M., Delgado-Battenfeld, C., and Jannach, D.: Beyond accuracy: Evaluating recommender systems by coverage and serendipity, Proc. ACM RecSys, pp.257-260 (2010)
- [26] Zhou, T., et al.: Solving the apparent diversity-accuracy dilemma of recommender systems, Proc. National Academy of Sciences, Vol.107, No.10, pp.4511-4515 (2010)
- [27] Bertsekas, D.P.: Nonlinear Programming Second Edition, Athena Scientific, Belmont, Massachusetts (1999)
- [28] Griffiths, T.L. and Steyvers, M.: Finding Scientific Topics, Proc. of National Academy of Sciences, Vol.101, pp.5228-5235 (2004)