4O2-OS-17b-2

多基準評価値に基づく協調フィルタリングと評価値統合

Collaborative Filtering and Rating Aggregation Based on Multicriteria Rating

森瀬 寛己 Hiroki Morise　小山 聡 Satoshi Oyama　栗原 正仁 Masahito Kurihara

北海道大学
Hokkaido University

In recent years, ratings by users on various items such as hotels and movies are easily available on the Web. In many cases, other than overall rating for each item by each user, more detailed information such as ratings from different viewpoints and free text comments, as well as aggregated information such as the average of ratings by different users, are also available. In this paper, we investigated the effectiveness of the existing collaborative filtering methods for large-scale sparse multicriteria rating data. We formulated rating aggregation as a collaborative filtering problem and also applied the collaborating filtering methods to this problem. Furthermore, we extended the existing methods by calculating user similarity using indirect users and review comments and applied them to collaborative filtering and rating aggregation.

1. はじめに

近年，レストランやホテル，映画やゲームなどに対する，利用者の評価が Web 上で容易に得られるようになっている．そこではあるユーザの一つのアイテム（対象）に対する数値による総合評価だけでなく，各観点での評価値や利用者のレビューコメントといったより詳細な情報や，複数の利用者の評価を統合した評価値も同時に提供されていることが多い．これらの情報を利用し，購買行動の補助をするものとして情報推薦システムがある．情報推薦システムの基本方式として協調フィルタリングがあるが，利用者の行動履歴と他の利用者の行動履歴から類似性を分析し，推薦を行う協調フィルタリングでは，スパースなデータに対して有効に作用しにくいという問題がある．

本研究ではホテルの立地，部屋，食事，風呂，サービス，設備に対する多基準評価値を含んだ大規模かつスパースなデータに対して，既存の協調フィルタリング手法の有効性を検証するとともに，複数の利用者からの評価値の統合への適用可能性を検証する．さらに，間接的なユーザ間の類似度と文章で書かれたレビューコメントも利用することで，より高精度な協調フィルタリングおよび評価値統合を実現する方法を検討する．

2. 協調フィルタリング

ユーザベースの協調フィルタリングとは，あるユーザがまだ評価していないアイテムに対する評価値を，類似のユーザの評価値から予測する方法である．

2.1 単一評価

従来の手法ではユーザがこれまでに付けた他のアイテムへの総合評価値からユーザ間の類似度を計算し，新規アイテムの評価値の予測を行う．類似度を測る一般的な方法としてコサイン類似度が用いられる．

類似度から対象ユーザの近傍集合 N を決定し，対象ユーザ u のアイテム i に対する評価の予測値 $\mathrm{pred}(u,i)$ は以下のように計算される．

$$\mathrm{pred}(u,i) = \overline{r_u} + \frac{\sum_{u' \in N} \mathrm{sim}(u,u')(r_{u'i} - \overline{r_u})}{\sum_{u \in N} |\mathrm{sim}(u,u')|}$$

ここで，$\overline{r_u}$ はユーザ u の評価値の平均を表す．

2.2 多基準評価値

従来の手法を多基準評価値を含むデータに適用する方法を説明する [Adomavicius 07]．ここではユーザがアイテムに対して総合評価と k 個の観点からなる項目に評価をしたとする (式 1)．

$$\mathrm{R}(u,i) = (r_0, r_1, \cdots, r_k) \tag{1}$$

ユーザは 1 つのアイテムに合計 $k+1$ 個の評価をしたことから，ユーザ同士の類似度の計算は $k+1$ 個できる．これらの類似度からユーザ同士の全体の類似度を計算し，従来の手法に適応する．ここでは，全体の類似度を各観点の類似度の平均値とする方法 (式 2) と最も類似度が低い観点の類似度を全体の類似度とする方法を示す (式 3)．

$$\mathrm{sim}_{\mathrm{avg}}(u,u') = \frac{1}{k+1} \sum_{i=0}^{k} \mathrm{sim}_i(u,u') \tag{2}$$

$$\mathrm{sim}_{\mathrm{min}}(u,u') = \mathrm{min}_{i \in \{0,\ldots,k\}} \mathrm{sim}_i(u,u') \tag{3}$$

次に距離を用いた類似度の計算方法を説明する．アイテムの各観点の評価値を $k+1$ 次元空間で考える．

距離の計算方法としてマンハッタン距離，ユークリッド距離，チェビシェフ距離 (式 4) がある．

$$\mathrm{max}_{i \in \{0,\ldots,k\}} |r_i - r'_i| \tag{4}$$

次に上記で得られたユーザ同士の距離 d_{user} から全体の類似度 (式 5) を計算する．距離が 0 になる場合を考慮し，分母に定数 1 を加える．

$$\mathrm{sim}(u,u') = \frac{1}{1 + d_{\mathrm{user}}(u,u')} \tag{5}$$

連絡先: 森瀬 寛己，北海道大学，
morise.hiroki@complex.ist.hokudai.ac.jp

表 1: 評価値集合

	アイテム 1	アイテム 2	アイテム 3	アイテム 4
ユーザ a	5	2	?	?
ユーザ b	?	?	2	3
ユーザ c	3	4	4	3

2.3 集約関数

ここでは，総合評価は各観点の評価とは独立しており，総合評価値は他の評価値とある関係 $r_0 = f(r_1, \cdots, r_k)$ によって成り立っているという考えからアイテムの評価値の予測を行う [Adomavicius 07].

Step1:各観点の評価値の予測

まず，予測を行うユーザのアイテムに関して，総合評価以外の各観点の評価値 $r_1, r_2, \cdots, r_k$ を予測する．

Step2:集約関数の学習

次に，アイテムの総合評価値と他の観点の評価値との関係 $r_0 = f(r_1, \cdots, r_k)$ を学習する．

Step3:総合評価値の予測

最後に予測した各観点の評価値と学習した集約関数を用いて，アイテムの総合評価値 $r'_0 = f'(r'_1, \cdots, r'_k)$ を予測する．

3. 評価値統合

複数のユーザのあるアイテムに対する評価値を統合する最も単純な方法は，そのアイテムを評価したユーザの評価値の平均をとる方法である．しかし，評価したユーザ数が少ない場合，特定のユーザの影響が大きくなり，ユーザの評価値の信頼性が低い場合，統合後の評価値の信頼性も低くなる問題がある．実際，評価者の数が少ないアイテムに対しては，統合評価値を表示しない Web サイトも多い．近年，クラウドソーシングによってワーカと呼ばれる不特定多数の作業者から得られた結果を統合する研究においては，ワーカの能力 [Dawid 79]，問題の難しさ [Whitehill 09]，ワーカの自信 [Oyama 13] などを考慮することで，信頼性の高いラベル統合を実現している．また，上記のような 2 クラスや多クラス分類の問題だけなく，一つのアイテムが複数のラベルを同時に持つマルチラベルデータや [Duan 14]，評価値データ [Uebersax 93] における結果統合の研究も行われている．評価値統合の従来研究は，単一評価値を前提とし，ワーカの能力や問題の性質を確率モデルでモデル化するというアプローチをとっている．これに対して本研究では，評価値統合を「平均的ユーザ」という仮想的なユーザに対する情報推薦の問題としてとらえ，多基準評価値データにおける協調フィルタリングアルゴリズムの有効性を検証する．

4. 拡張手法

データのスパース性によりユーザ同士の類似性が計算できないという問題を解消するために，間接的なユーザ間の類似度関係を用いる方法，ユーザのレビューコメントを用いる方法の 2 種類の拡張を行う．

間接的なユーザを利用する方法では，例えばユーザ a とユーザ b は共通して評価しているアイテムが無いため類似度を計算することができないが，ユーザ a とユーザ c，ユーザ b とユーザ c では共通して評価しているアイテムが存在するため，類似度を計算できるような場合があるとする (表 1)．このとき，ユーザ a とユーザ b の類似度 $\mathrm{sim}(a, b)$ はユーザ a とユーザ c の類似度 $\mathrm{sim}(a, c)$ とユーザ b とユーザ c の類似度 $\mathrm{sim}(b, c)$ から以下のように計算する．

$$\mathrm{sim}(a, b) = \frac{(\mathrm{sim}(a, c) + \mathrm{sim}(b, c))}{2}$$

間接的なユーザが複数いる場合は，間接的なユーザ集合 N より全体の平均値をとる (式 6).

$$\mathrm{sim}(a, b) = \frac{\sum_{u \in N} \mathrm{sim}(a, u) + \mathrm{sim}(b, u)}{2|N|} \quad (6)$$

また，ユーザ同士で共通して評価しているアイテムが無い場合，レビューコメント中の語に基づくベクトルで，ユーザ同士の類似度を計算する．まず，ユーザが投稿した全てのレビューコメントをまとめて 1 つの文章 D_u とする．次に形態素解析で全ての品詞を取り出し，TF-IDF 法で重みを付けた特徴ベクトル D'_u を用いて，コサイン類似度 $\mathrm{sim}(D'_u, D'_{u'})$ を計算する．レビューコメントを用いた場合，共通して評価しているアイテムが無くても，レビューコメントは全てのユーザで集められるため，全てのユーザ間の類似度を計算することができる．

5. 実験

協調フィルタリングと評価値統合について実験を行った．

5.1 協調フィルタリング

5.1.1 実験設定

既存の協調フィルタリング手法と拡張手法の比較実験を行った．本研究では多基準評価を含んだ大規模かつスパースなデータセットとして楽天データセットの楽天トラベルデータを使用した．楽天トラベルデータには施設データとレビューコメント，多基準評価値 (立地，部屋，食事，風呂，サービス，設備，総合) が記載されている．評価値は 1 点から 5 点の 5 種類を採用する．総データ数はユーザ 881 人、ホテル 5098 個，評価数 16993 件である．各ユーザに関して，総合評価値の違うアイテムを 2 つランダムに抽出し，それらのアイテムを未評価なアイテムと仮定する．未評価と仮定した全てのアイテムをテストデータとし，評価値の予測を行う．

予測結果の評価方法として，各ユーザで予測した 2 つのアイテムの評価値の大小関係が合っていれば正解，間違っていれば不正解とし，全体の正解率を計算する．テストデータを変え，各手法 100 回ずつ予測を行った結果を次に示す．

5.1.2 結果と考察

表 2 に結果を示す．

以下に各手法について簡単に説明する．

Standard CF

総合評価値からコサイン類似度を計算する手法，従来の手法で本研究のベースラインとなる．

Cos-min

総合評価を含めた，各観点の評価値からコサイン類似度

表 2: 実験結果 (正解率)

手法＼拡張点	なし	間接的ユーザ	レビューコメント
Standard CF	58.1	60.7	62.7
Cos-min	58.1	60.8	62.7
Chebyshev	59.3	62.2	62.7
Aggregation	61.3	62.0	63.4

を計算し，個々の類似度の最小類似度を全体の類似度とする手法.

Chebyshev
総合評価を含めた，各観点の評価値のチェビシェフ距離から類似度を求める手法.

Aggregation-total
集約関数を用いた手法，総合評価値を考慮した従来の協調フィルタリングで各観点の評価値を予測し，学習した線形回帰モデルから総合評価値を予測する手法.

まず拡張点による正解率の変化を見ると，間接的なユーザ，レビューコメントを用いた場合で精度の向上が見られた．次に各既存手法ごとの正解率を見ると，チェビシェフ距離，集約関数を用いる手法での有効性が見られた．一方で，レビューコメントを利用した場合では，総合評価値のみを考慮する手法，コサイン類似度の最小値をとる手法，チェビシェフ距離を用いる手法で正解率の差が見られなかった．これは各手法で類似度を計算するよりも，レビューコメントで類似度を計算したものの影響が大きいことが原因と考えられる．

5.2 評価値統合

5.2.1 実験設定

楽天トラベルデータセットの中から 15 種類以上のホテルを評価しているユーザ，15 人以上のユーザから評価されているホテルを抽出し，新たなデータセットを作成する．各アイテムを評価したユーザの総合評価値の平均値を統合評価値とする．これらのアイテムは多くのユーザから評価を受けているため，このようにして計算した統合評価値は信頼性が高いと仮定する．このようにして得られたデータを訓練データとテストデータに分割し，訓練データでのユーザの評価値と各アイテムの統合評価との類似性からテストデータのアイテムの統合評価を予測する．予測するアイテムに評価をしているユーザの数，予測を行う手法を変え実験を行う．

評価方法としてこの実験では 2 乗平均平方根誤差 (RMSE) を用い，5 分割交差検証を用いた実験結果を次に示す.

5.2.2 結果と考察

表 3 に結果を示す．レビューコメントを利用する手法としてチェビシェフ距離での類似度とレビューコメントでの類似度の平均の類似度を用いる手法 Chebyshev+review を追加した．

評価値統合についても総合評価値のみを考慮する従来の手法と比較して，多基準評価を考慮する手法で同等，もしくは良い結果が得られた．一方で集約関数を用いた手法では既知ユーザ数少ない場合は結果が良く，有効に作用していたが，既知ユーザ数が増えたときには他の手法のほうが良い結果となった．これは学習した線形回帰モデルが関係していると考えられる．学習データは既知ユーザ数に関わらずほぼ同じであるため，既知ユーザ数の増加によって各観点の予測値の精度が上がったとしても，有効に作用しなかったのではないかと考える．

表 3: 実験結果 (RMSE)

既知ユーザ数 / 手法	2	4	6	8	10
Standard CF	0.45	0.35	0.26	0.19	0.16
Cos-min	0.45	0.35	0.26	0.19	0.16
Chebyshev	0.45	0.33	0.24	0.16	0.12
Aggregation-total	0.46	0.32	0.26	0.22	0.19
Chebyshev+review	0.45	0.32	0.23	0.16	0.11

6. まとめ

6.1 結論

本研究では多基準評価値を含んだ大規模かつスパースなデータに対して，2 つの実験を行った．協調フィルタリングでは拡張手法における精度の向上，チェビシェフ距離と集約関数を用いる手法の有効性を示した．評価値統合では既存手法においてチェビシェフ距離を用いる手法での有効性を示した.

6.2 今後の展望

本研究ではユーザベースの協調フィルタリング手法を用いたが，欠損値補完や行列分解などのモデルベースの手法 [Adomavicius 15] もあり，これらの手法を多基準データにおける情報推薦や評価値統合に適用する研究を行いたい.

謝辞

本研究の作成に当たり，楽天株式会社より楽天トラベルにおける施設データ及びレビューデータを使用させていただいた．また，本研究の一部は JSPS 科研費 15H0278206 の支援を受けた．この場を借りて深く感謝する.

参考文献

[Adomavicius 07] Adomavicius, G. and Kwon, Y.: New recommendation techniques for multicriteria rating systems, *IEEE Intelligent Systems*, Vol. 22, No. 3, pp. 48–55 (2007)

[Adomavicius 15] Adomavicius, G. and Kwon, Y.: Multicriteria recommender systems, in *Recommender systems handbook*, pp. 847–880, Springer (2015)

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum likelihood estimation of observer error-rates using the EM algorithm, *Applied statistics*, pp. 20–28 (1979)

[Duan 14] Duan, L., Oyama, S., Sato, H., and Kurihara, M.: Separate or joint? estimation of multiple labels from crowdsourced annotations, *Expert Systems with Applications*, Vol. 41, No. 13, pp. 5723–5732 (2014)

[Oyama 13] Oyama, S., Baba, Y., Sakurai, Y., and Kashima, H.: Accurate integration of crowdsourced labels using workers' self-reported confidence scores., in *IJCAI*, pp. 2554–2560 (2013)

[Uebersax 93] Uebersax, J. S. and Grove, W. M.: A latent trait finite mixture model for the analysis of rating agreement, *Biometrics*, pp. 823–835 (1993)

[Whitehill 09] Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., and Ruvolo, P. L.: Whose vote should count more: Optimal integration of labels from labelers of unknown expertise, in *NIPS*, pp. 2035–2043 (2009)