情報処理学会論文誌 Vol.56 No.1 97–105 (Jan. 2015)

行列因子分解によるWebユーザビリティ評価値の予測

山田 俊哉[1,a)] 中道 上[2] 松井 知子[3]

受付日 2014年4月13日, 採録日 2014年10月8日

概要：Webサイトのユーザビリティテスティングにおいて，被験者の評価傾向に適したWebページを評価対象として選択することによって，その被験者はより多くのユーザビリティ問題を指摘できる可能性があるが，そのようなWebページをタスクとして推薦することは難しい．本研究では，被験者が評価を完了させた一部のWebページの評価結果から被験者の評価傾向を分析し，その傾向に適した評価対象Webページを推薦するため，行列因子分解を用いて被験者が未閲覧のWebページの評価値を予測することが可能か検証する．行列因子分解は (Webページ数) × (被験者数) で表す評価値行列をユーザの潜在因子，Webページの潜在因子に分解する手法である．基本的な行列因子分解に加え，バイアス付き行列因子分解と，評価値の重み付き行列因子分解を用いた予測法を試みた．実際にユーザビリティテスティングを実施し，収集したユーザビリティ評価値行列は84%欠損値を含んでいた．この評価値行列に対し行列因子分解による予測法を適用した結果，重み付き行列因子分解では，ユーザビリティの4段階評価において1段階以下の誤差で評価値の予測が可能であった．

キーワード：Webユーザビリティ，ユーザビリティテスティング，協調フィルタリング，行列因子分解

Score Prediction in Web Usability Evaluation Using Matrix Factorization

Toshiya Yamada[1,a)] Noboru Nakamichi[2] Tomoko Matsui[3]

Received: April 13, 2014, Accepted: October 8, 2014

Abstract: In usability testing of websites, there are web pages which only a small number of subjects evaluated. It is difficult for evaluators to determine whether to include such a page in the redesign. We predict a value when other subjects evaluate, and support determination. It is verified whether the evaluation value of the Web page which is not visited can be predicted using matrix factorization. Usability evaluation values were collected by usability testing. The missing value was included in 84% of usability evaluation value matrix. The predicting method by matrix factorization was applied to the evaluation value matrix. Weighted matrix factorization has the highest predictive accuracy as analysis results. In 4-point scale evaluation of Web usability, the range of the error by it was one or less point.

Keywords: Web usability, usability testing, collaborative filtering, matrix factorization

1. はじめに

Webはネットワークサービスを提供する窓口として，また知識や情報の共有・発信するためのICT環境を提供する基盤として，ますますその重要性が高まっている．企業においてもWebサイトの事業貢献度は高まっており，その中で数百億円を超える経済価値を持つWebサイトも存在する [15]．そのため，Webユーザビリティは重要性が高く，使いやすいWebサイトを構築するため，反復デザイン [13] が実施され，デザインのバージョンは次々に検討されている．その際，各々のバージョンに対しては，ユーザビリティテスティング [11], [12] のようなユーザビリティ評価を実施し，これらのユーザビリティ上の発見を基に次

1 NTTアイティ株式会社
NTT IT CORPORATION, Yokohama, Kanagawa 231–0032, Japan
2 福山大学工学部
Fukuyama University, Fukuyama, Hiroshima 729–0292, Japan
3 情報・システム研究機構統計数理研究所
The Institute of Statistical Mathematics, Tachikawa, Tokyo 190–8562, Japan
a) yamada.toshiya@ntt-it.co.jp

のバージョンが修正される．ユーザビリティテスティングは被験者を集めるコストが大きいが [2]，実際に使う場面でのユーザビリティの問題点を発見できるため広く用いられている．

Web サイトに対するユーザビリティテスティングは，実際に Web サイトを被験者に閲覧してもらい，その Web サイトの中で閲覧した Web ページを被験者自身に評価をしてもらう方法である．実際に被験者が評価するため対象となる Web サイトの閲覧目的を具体的にする必要があり，被験者には Web サイトから特定の情報を発見するといったタスクを課する [6]．Web ページごとの主観的満足度を被験者評価してもらう際，評定加算法などの評価点数を用いて短く単純な質問を行うことが多い．Web ページごとの主観的満足度の評価後，タスク実施時の様子を再生しつつ評価した Web ページごとにユーザビリティに関するインタビュを実施する．その結果，改善が必要と思われる Web ページを抽出し，問題個所を抽出する．

評価対象の Web サイトを構成する Web ページ数が多い場合，タスクの達成までに被験者が閲覧する Web ページの順序，種類は被験者ごとに異なる．1 人の被験者がすべての評価対象となる Web ページを評価することは難しいため，(Web ページ) × (被験者) の評価値行列には欠損値が多く含まれている．このように少数の被験者しか評価していない Web ページがある場合や，ごく少数の被験者が偏った評価傾向を持つ場合がある．このような場合，次のバージョンの改善に向けての修正対象とするべきか判断することが困難である．そのため修正対象としての優先度は低くなり，該当ページの評価値は有効に利用されない場合がある．

リクルーティングした被験者に複数のタスクを実施する場合や，早くテストが終わった被験者に追加タスクを与えるなどがユーザビリティテストの現場ではなされている．一方で被験者には評価に一定の傾向が存在する場合があり，被験者ごとに指摘可能なユーザビリティ問題は異なる．こうした傾向をすでに終了した一部のタスクから推定することによって，追加的にタスクを選択する際に被験者の評価傾向に基づいてタスクを選定することが可能となり，より多くの問題発見の効果が期待される．

本論文では，Web サイトに対するユーザビリティテスティングにおいて，被験者が評価完了した一部の Web ページの評価を用いて被験者の評価傾向を推定し，追加的に実施する被験者評価のタスク選定時に評価傾向に合ったタスク選定を行うことを目的とする．そのため，欠損値を予測し，ある被験者が未閲覧の Web ページの評価値を推定し，その被験者に未閲覧の Web ページの中で被験者がユーザビリティの問題を発見しやすいであろうページを推薦する．具体的には，欠損値を多く含む (Web ページ) × (被験者) の評価値行列に対し，協調フィルタリング手法の 1 つである行列因子分解を用いて，被験者が未閲覧の Web ページの評価値を推定する．これにより，限られた Web ページしか評価していない被験者であっても，予測された評価傾向を合わせて用いることによって，被験者が問題発見しやすいであろう Web ページを次のタスクとして提示することが可能となる．

本論文では 2 章で本論文における Web ユーザビリティの定義について説明する．3 章では行列因子分析を用いた評価値行列の予測法について述べる．4 章では行列因子分解による評価値行列の予測法の適用実験について述べ，5 章では考察，6 章ではまとめと今後の展望について述べる．

2. Web ユーザビリティ

ここでは本論文における Web ユーザビリティについて述べる．一般的に Web ページ閲覧行動は見るだけではなく，操作も含まれる行動である．そのため，Web ユーザビリティはソフトウェアユーザビリティと同様に考え，「見やすさ」ではなく「使いやすさ」として考えることができる [12], [14]．Nielsen は，ユーザビリティの重要な要因の 1 つとして，ユーザによる主観的満足度をあげている [12]．Web サイトでは，通常の製品とは異なり，ユーザの主観的満足度が低い場合，ユーザは使い続けるより，他の Web サイトの利用を選択する傾向がある．そのため，Web サイトはユーザの主観的満足度を低下させないように設計される必要がある．ユーザの主観的満足度に関する評価値を測るため，Web サイトを閲覧するユーザを被験者としユーザビリティテスティングが行われている．

本論文では，Web ページに対するユーザビリティテスティングにおける被験者の主観的満足度に関する評価値を分析対象とし，Web ページ単位でユーザビリティ問題の改善を行うことを想定し，その評価値を 1 ページ単位で得るものとする．またユーザビリティテスティングを行う際，被験者がタスク遂行途中で同じ Web ページを複数回閲覧する場合がある．複数回閲覧したページは，それぞれ別のページとして扱い，被験者が閲覧した Web ページの数を PV（Page View）としてカウントし評価値を得る．

3. 行列因子分解による予測

一般的にユーザ（本論文では被験者に相当）による様々なアイテム（本論文では評価値に相当）を測る際，あるユーザの評価値を事前に嗜好を取集したユーザの嗜好を用いて推定する手法として，協調フィルタリングが広く用いられている [3]．協調フィルタリングは主に Amazon [1] や Netflix [9] などにおける商品推進システムに用いられる方法である [4]．従来，協調フィルタリングの手法では，あるユーザの嗜好を推定する GroupLens 法などの方法が用いられている [17], [18]．

また，協調フィルタリングにおいてユーザ間の類似度を

測る方法以外に，ユーザによるアイテムに対する嗜好や評価を並べた評価値行列を，ユーザの潜在因子，アイテムの潜在因子に分解する手法である行列因子分解が注目されている．この手法では評価値行列の分解のみに基づくため，直接ユーザ間の類似度を求める必要がない．さらに，それぞれの潜在因子より嗜好の推定を行うことができる．行列因子分解による協調フィルタリングは，推薦システムに関する国際的なコンペティション Netflix Prize competition [10] においてトップの成績を収めた手法である [7].

本論文では，協調フィルタリング手法の 1 つである行列因子分解を用いて Web ページのユーザビリティテスティングにおける評価値の予測を行う．行列因子分解は評価値を，被験者の評価傾向に関するベクトルと Web ページの評価される傾向に関するベクトルに分解し，再び結合することで予測する方法である．本章では行列因子分解の説明と，行列因子分解による評価値行列の予測法について述べる．

3.1 基本的な行列因子分解

行列因子分解では，被験者 u による，Web ページ i の評価値 r_{ui} を Web ページ潜在因子ベクトル $q_i \in R_f$ とユーザ潜在因子ベクトル $p_u \in R_f$ の内積で表す．行列因子分解において，評価値 r_{ui} は式 (1) で表される．

$$r_{ui} = q_i^T p_u \tag{1}$$

1 章で述べたように，Web ページに対するユーザビリティテスティングにおいては，被験者が評価対象の Web ページを一部しか評価できず，評価値行列 r_{ui} は多くの部分が欠損値となる．欠損値を多く含む行列において，既知の r_{ui} に対し，式 (2) を用いて q_i，p_u を推定する．

$$\min_{q,p} \sum_{(u,i)\in\kappa} \left(r_{ui} - q_i^T p_u\right)^2 + \lambda\left(\|q_i\|^2 + \|p_u\|^2\right) \tag{2}$$

κ は入力データ中に存在する (u,i) のすべてのペアであり，λ は，正則化パラメータである．本論文では，式 (2) のメリット関数を用いる行列因子分解を「基本分解」と呼称する．

ここで，この最適化問題は凸計画問題ではない．しかし，交互最小 2 乗法（alternating least squares; ALS）により q_i，p_u をもとめることが可能である [5]．これは q_i，p_u のどちらかのベクトルを固定した場合，固定していない方のベクトルに対しては 2 次計画問題となっていることを利用し，交互に最小 2 乗法を適用して繰り返し解く方法である．本論文では ALS により q_i，p_u を推定する．

3.2 バイアス付き行列因子分解

特定の Web ページに対し評価値を低くつける傾向など，被験者や Web ページ間で評価の揺らぎが存在することはよく知られており，これをバイアス項として行列因子分解に用いる [16]．そこで被験者の評価バイアス項 b_u と，Web ページの評価バイアス項 b_i を前述の基本的な行列因子分解に追加する．r_{ui} の平均値を μ としたとき，評価値 r_{ui} の推定値は式 (3) のように表される．

$$r_{ui} = \mu + b_u + b_i + q_i^T p_u \tag{3}$$

ここで μ は入力データにおける欠損値ではないすべての評価値の平均値である．バイアス項を考慮した場合，式 (4) を用いて q_i，p_u，b_i，b_u を推定する．

$$\min_{q,p} \sum_{(u,i)\in\kappa} \left(r_{ui} - \mu - b_u - b_i - q_i^T p_u\right)^2 + \lambda\left(\|q_i\|^2 + \|p_u\|^2 + b_u^2 + b_i^2\right) \tag{4}$$

この問題にも同様に ALS を用いることができる．本論文では式 (4) のメリット関数を用いた行列因子分解を「バイアス付き分解」と呼称する．

3.3 評価値の重み付き行列因子分解

本節では既知の評価値に対しその重みを考え，Web ページに対するユーザビリティ評価値の予測に利用した方法を提案する．ユーザビリティテスティングの実施目的はユーザビリティの低い Web ページの発見および問題点の分析である．そのため，著しく評価値が平均値から外れたページに関してはユーザビリティ問題の分析が必要であると考えられ，このような評価値を重視する必要がある．ユーザビリティテスティング評価値に適用させるため，各評価値に対する重み c_{ui} を導入する．重みを考慮した行列因子分解では，式 (5) を用い q_i，p_u，b_i，b_u を推定する．

$$\min_{q,p} \sum_{(u,i)\in\kappa} c_{ui}\left(r_{ui} - \mu - b_u - b_i - q_i^T p_u\right)^2 + \lambda\left(\|q_i\|^2 + \|p_u\|^2 + b_u^2 + b_i^2\right) \tag{5}$$

ここで重みは，平均値から離れた評価値には大きく，平均値に近い評価値には小さくなるよう設定し，平均値から離れた評価値を精度良く推定することを試みる．そのため重み c_{ui} は，次の重み A，B，C の 3 つの種類で表し，それぞれ以下のようになる．

重み A：全評価値の平均値 μ からそれぞれの r_{ui} のユークリッド距離

$$c_{ui} = (\mu - r_{ui})^2 \tag{6}$$

重み B：被験者ごとの評価値の平均値 μ_u からの距離

$$c_{ui} = (\mu_u - r_{ui})^2 \tag{7}$$

重み C：Web ページごとの評価値の平均値 μ_i からの距離

$$c_{ui} = (\mu_i - r_{ui})^2 \tag{8}$$

それぞれの重み c_{ui} が大きいほど，それぞれの平均値からの差の二乗が大きい．そのため，被験者は他の Web ページに対し著しい評価を付けていると考えられ，重みが大きい評価値をより誤差が小さくなるように分解できる．本論文では式 (5) のメリット関数を用いた行列分解において，重みに全評価値の平均値からの差の二乗を用いた行列因子分解を「重み付き分解 A」，被験者ごとの評価値の平均値からの差の二乗を用いた分解を「重み付き分解 B」，Web ページごとの評価値の平均値からの差の二乗を用いた分解を「重み付き分解 C」と呼称する．

4. 評価値行列の予測実験

本章では行列因子分解を Web ページのユーザビリティ評価に適用した評価値行列の予測実験について述べる．はじめに，予測実験に用いたユーザビリティ評価値を収集するために実施したユーザビリティテスティングの実施環境について述べ，次にユーザビリティテスティングの実施方法，被験者および評価対象の Web ページについて述べ，本論文で用いる評価値データについての概要を述べる．最後に行列因子分解によるユーザビリティ評価値行列の予測結果について述べる．

4.1 ユーザビリティ評価値データの収集

予測実験には，実際のユーザビリティテスティングにおいて収集した Web ページのユーザビリティ評価値データを用いた．本節ではユーザビリティ評価値データを収集するためのユーザビリティテスティングについて，実施環境，被験者とタスク，そして実施手順と収集されたデータについて説明する．

4.1.1 ユーザビリティテスティング実施環境

ユーザビリティテスティングの実施環境は以下のとおりである．

- ディスプレイ：液晶 21 インチ（有効表示領域：縦 30 cm，横 40 cm，解像度：1280 pixel × 1024 pixel）
- 顔とディスプレイの距離：約 50 cm
- Web ブラウザ：Firefox 3.6.6
- Web 閲覧行動記録：ITR-Recorder [8]
- Web 閲覧行動再生：ITR-Player [8]

ITR-Recorder/Player は，Web ページ閲覧中の被験者のブラウザ操作を記録するツールであり，各被験者の Web ページ閲覧時のブラウザ画面を再生することが可能な Firefox のアドオンツールである [8]．

4.1.2 被験者とタスク

被験者は，日常的にインターネットを利用している理工系の大学生および大学院生 35 名である．ユーザビリティテスティング実施時点で，被験者は，実験対象に設定した 8 つの企業の Web サイトを閲覧した経験はない．

まず，予備実験として，被験者にユーザビリティテスティング実施環境に慣れてもらうことを目的に，あるポータルサイトからニュースを 2 つ読むタスクを行うよう依頼し，日常使用している Web ページ閲覧環境と比べ，ユーザビリティテスティング実施環境に大きな違和感を覚えないか確認した．

次に，本実験として，指定した企業の Web サイトから大学卒者の初任給を探すというタスクを行うよう依頼した．タスクで使用する企業サイトは，これらの 8 つの企業サイトのうち，3 から 8 つの企業を無作為に指定しタスクを実行するよう指定した．1 人の被験者が行うタスク数は最小で 3 タスク，最大で 8 タスクであり，それぞれの被験者ごとのタスクの実行順序はランダムに決定した．タスクの開始は指定された企業サイトの Top ページからであり，タスクの達成は被験者が大卒者の初任給についての情報を発見した時点とした．

タスク全体に対する速さや慣れが評価に影響することを少なくし，Web ページそのもののユーザビリティを被験者に評価してもらうため，それぞれのタスク実行後に，操作履歴を再生し，それを被験者に閲覧してもらい，タスク実行時に閲覧した 1 PV ごとに被験者のユーザビリティ評価値を得た．

4.1.3 ユーザビリティテスティング実施手順

ユーザビリティテスティングの実施手順を以下に示す．

手順 1：初期設定として，被験者のディスプレイに各企業のトップページへのリンクを張った実験用 Web ページを表示しておき，タスクを実行するために被験者がそのリンクをクリックした時点から実験を開始する．

手順 2：被験者のタスク実行中のブラウザ操作の様子を ITR-Recorder を用いて記録する．その際，評価者が被験者に対して質問するといったタスクの中断につながることは行わなかった．タスクは被験者が初任給を見つけることができたと申告した時点で終了する．

手順 3：被験者が感じる評価値を収集するため，タスク終了後すぐに ITR-Player を用いて被験者の Web ページ閲覧記録を再生しながら，Web ページ 1 PV ごとの使いやすさを下記の 4 段階から選択するよう依頼する．

評価値 = 1：使いにくい
評価値 = 2：どちらかといえば使いにくい
評価値 = 3：どちらかといえば使いやすい
評価値 = 4：使いやすい

本論文では，上記の評価に対応した評価値を用いる．評価の際には複数の Web ページにまたがる印象評価を避け単一の Web ページとして評価するように依頼した．これはある Web ページ A において目的とは異なるリンクをたどりあるページ B に遷移した後被験者が目的と異なるページであると気づきページ A に戻った場合，被験者は遷移した順番に沿って Web ページ A は 2 回，Web ページ B は 1 回評価する．この際に最初の Web ページ A の評価は誤っ

てページ B に誘導する要因を含めて評価し Web ページ B では目的と異なると気付けたか，戻るためのインタラクションがすぐできたかなどを含めて評価するように依頼した．本論文では被験者 u が Web ページ i を「使いにくい」と評価した場合そのページの評価値は $r_{ui} = 1$ として扱う．また，(Web ページ数) × (ユーザ数) の評価値行列を作成するため，Web ページ 1 ページに対し 1 ユーザがつける評価値を 1 PV 分に集約する必要がある．そのため 1 人の被験者が同じ Web ページを複数回閲覧した場合，閲覧 PV 数分の評価値の最小値をその Web ページの評価値として集約する．これは，先の例で示した Web ページ A のような場合，2 回目に遷移した際の評価は Web ページ B への遷移が誤りであると理解したうえでの評価となるため，この場合でも Web ユーザビリティの評価において使いにくいとされるページを重視して分析を行うためである．

4.2 収集された評価値データ

被験者 35 名によるユーザビリティテスティングの実施により，ユーザビリティ評価を収集した Web ページは 93 ページである．複数回閲覧された Web ページの評価値を集約すると，評価値は 498 PV であり，そのうち「使いにくい」= 59 PV,「どちらかといえば使いにくい」= 113 PV,「どちらかといえば使いやすい」= 145 PV,「使いやすい」= 181 PV である．

被験者 35 名，93 ページにおいて 498 PV 分の評価値であるため，93 行 × 35 列の評価値行列のうち，498 カ所に評価値が存在し，残りの 2,757 カ所が欠損値となる．つまり 84%の部分で評価値が欠損した状態である．

4.3 行列因子分解による Web ユーザビリティ評価値の予測結果

収集されたユーザビリティ評価値行列に対し，基本分解，バイアス付き分解，重み付き分解 A，重み付き分解 B，重み付き分解 C の 5 つの予測法と，特異値分解（SVD）および GroupLens 法により予測を試みた．

まずそれぞれの予測法において，パラメータを決定するため予備実験を行う．予備実験では予測法のパラメータである λ を $0.001 \leqq \lambda \leqq 2.0$ まで変化させ，学習データとテストデータの両方の RMSE を計算する．予備実験に用いるデータは，評価値データより，評価値の存在するデータから 2/3 を無作為抽出して学習データとし，残りをテストデータとした．この学習データとテストデータの組を 5 個作成し，学習データ，テストデータの平均 RMSE を計算した．

ユーザビリティテスティングにより得られた評価値行列の一部を**表 1** に表し，**表 2** に上記評価値行列を予測した結果をあげる．表 1 はユーザビリティテスティングによって得られた評価値を行方向に Web ページ，列方向に被験者をとる評価値行列の一部である．表 2 は行列因子分解による評価値行列の予測後の評価値行列の一部であり，欠損値のない行列となっている．行列因子分解における潜在因子の次元数は 10 であり，パラメータ λ は予備実験より学習データに対する RMSE が最も小さい $\lambda = 0.1$ とした．

表 1 予測前の評価値行列（一部），“-” は欠損値
Table 1 A part of the evaluation matrix before prediction (“-”; missing value).

Web ページ	被験者					
	A	B	C	D	E	F
Page1	3.00	4.00	-	3.00	2.00	-
Page2	1.00	4.00	-	-	-	-
Page3	1.00	4.00	-	4.00	3.00	-
Page4	1.00	3.00	-	3.00	-	-
Page5	4.00	4.00	-	4.00	-	-
Page6	-	2.00	2.00	4.00	2.00	3.00
Page7	-	4.00	3.00	4.00	4.00	3.00

表 2 予測後の評価値行列（一部），太字は予測した値
Table 2 A part of the evaluation matrix after prediction (**boldface**; complemented value).

Web ページ	被験者					
	A	B	C	D	E	F
Page1	2.96	3.99	**2.45**	3.00	2.00	**3.47**
Page2	1.00	4.00	**2.81**	**2.94**	**3.02**	**3.24**
Page3	1.00	4.00	**3.02**	4.00	2.98	**3.05**
Page4	1.00	2.98	**3.11**	3.02	**3.02**	**2.35**
Page5	4.00	4.00	**2.85**	4.00	**3.37**	**3.32**
Page6	**1.89**	2.01	2.00	4.00	2.00	3.00
Page7	**3.08**	4.00	3.04	4.00	4.00	3.03

次に予測した評価値の精度を測るため，予備実験同様に得られた評価値データを学習データとテストデータに分け特異値分解および GroupLens 法と行列因子分解を用いた予測手法との比較を行った．行列因子分解により学習データの評価値行列を予測し，その予測した評価値行列とテストデータの評価値を比較し予測精度を測った．

学習データには評価値行列の中の欠損値ではない評価値の 2/3 にあたる 332 PV 分の評価値をランダムに抽出したものを用い，残りの 166 PV をテストデータとした．評価値データの中には，評価値をつけた被験者が 3 名以下のページに対する評価値が 74 PV 含まれる．そのため少数ユーザしか閲覧していないページのみでテストデータまたは学習データが構成されることを防ぐため，全評価値の 2/3 を学習データとしている．この学習データとテストデータの組を 100 セット作成し，それぞれの学習データについて評価値の予測を行い，テストデータを用いて精度を測った．

また本論文では精度を測るため平均二乗誤差（Root Mean Square Error; RMSE）を用いた．これはテストデータに存在する評価値と，その評価値に対応する予測後の評価値の平均二乗誤差であり，式 (9) で定義する．

表 3 100 組のデータセットの平均 RMSE
Table 3 Mean RMSE values for 100 data sets.

予測法	全評価	評価 1	評価 2	評価 3	評価 4
特異値分解	1.06	1.87	0.98	0.48	1.08
GroupLens 法	1.02	1.70	1.94	0.45	1.11
基本分解	1.45	1.23	1.14	1.49	1.62
バイアス付き分解	1.07	1.82	1.00	0.57	1.09
重み付き分解 A	0.91	1.43	0.86	0.45	1.00
重み付き分解 B	0.94	1.54	0.87	0.42	1.02
重み付き分解 C	0.98	1.60	0.88	0.38	1.10

表 4 100 組のデータセットの平均評価値
Table 4 Mean Score for 100 data sets.

予測法	平均評価値(予測前)	平均評価値(予測後)
特異値分解	2.90	2.90
GroupLens 法	2.90	2.87
基本分解	2.90	1.52
バイアス付き分解	2.90	2.85
重み付き分解 A	2.90	2.75
重み付き分解 B	2.90	2.78
重み付き分解 C	2.90	2.84

$$\sqrt{\frac{1}{n}\sum_{(i,u)\in\kappa}\left(r_{ui}-\hat{r}_{ui}\right)^2} \tag{9}$$

それぞれのデータセットの組で RMSE を計算し，100 セットの平均 RMSE を用いて各予測法を比較する各モデルにおいてパラメータ λ は予備実験においてテストデータに対する RMSE が最も小さいパラメータより決定し，基本分解：$\lambda = 0.4$，バイアス付き分解：$\lambda = 0.4$，重み付き分解 A = 重み付き分解 B = 重み付き分解 C：$\lambda = 0.85$ とした．また潜在因子の次元数は 10 とした．各モデルを用いた 100 組のデータセットの平均 RMSE は**表 3** のようになる．

ここで表 3 において，“全評価” の列はテストデータにおける全評価値に対する平均 RMSE を表し，“評価値 i, $(i = 1, \cdots, 4)$” の列は，テストデータにおけるそれぞれの評価値 i での平均 RMSE を表す．

以上の結果より，特異値分解と GroupLens 法と比較しても重み付き分解 A の精度が高い．t 検定により全評価値の RMSE の差を確認すると，平均 RMSE は (重み付き分解 A) < (重み付き分解 B) = (重み付き分解 C) < (バイアス付き分解) < (GroupLens 法) < (特異値分解) < (基本分解) の順となった．以上の結果より本研究における評価値データでは重み付き分解 A の予測精度が最も高いことが分かった．

一方で，予測前の評価値行列の平均評価値とそれぞれの手法を用いて予測した評価値行列の平均評価値は**表 4** のようになる．それぞれの手法における予測後の平均評価値の差に対して t 検定を行うと，すべての手法の組合せにおいて有意水準 5%で有意差が確認された．

表 5 100 組のテストデータにおける差の平均値
Table 5 Difference in evaluate values between practical testing and complementation.

予測法	全評価	評価 1	評価 2	評価 3	評価 4
特異値分解	-0.05	-1.82	-0.85	0.06	0.96
GroupLens 法	0.01	-1.65	-0.84	0.08	1.02
基本分解	-0.66	0.59	-0.07	-0.83	-1.30
バイアス付き分解	0.33	1.72	0.82	-0.06	-0.95
重み付き分解 A	-0.03	1.37	0.72	-0.11	-0.90
重み付き分解 B	-0.03	1.49	0.76	-0.11	-0.94
重み付き分解 C	-0.04	1.57	0.81	-0.12	-1.04

表 4 より，予測前と予測後の平均評価値の差が最も小さいのは特異値分解である．一方で，基本分解は平均評価値を低く予測する傾向がある．また重み付き分解 A は RMSE が最も小さく予測精度は高いが，予測前に比べて平均評価値は基本分解に次いで予測前と予測後の平均評価値の差が大きい結果となった．

ユーザビリティテスティングでは「使いにくい」ページの評価値を「使いやすい」と推定することを減少させることが重要である．そこで，RMSE ではなく，テストデータにおける被験者による実際の評価値と推定した評価値（予測した評価値）の差を確認する．100 組のデータセットでの差の平均値を**表 5** に示す．表 5 より各行列因子分解手法で評価値 1，評価値 2 に関してはテストデータにおける正解の評価値よりも大きく，評価値 4 では小さく推定されていることが分かる．また，基本分解では評価値 1 に関してテストデータの評価値より大きく推定するケースは少ない．しかし，全体的に評価値を小さく付ける傾向が強く，多くのページを「どちらかといえば使いにくい」という評価とする傾向があり，重み付き分解 A などに比べて予測の精度は低い．また，重み付き分解 A ではテストデータにおける評価値 1 に対応する評価値は，予測行列で平均 2.37 と推定されるため，「使いにくい」から「どちらかといえば使いにくい」の評価範囲に収まると考えられる．

実際に，重み付き分解 A を適用した例をあげる．あるデータセットにおいて，テストデータの実評価値と予測後の推定評価値（予測後の評価値）の関係を Box プロットにより**図 1** に表す．図 1 において，横軸はテストデータにおける正解の評価値を表し，縦軸は予測後の評価値行列の評価値を表す．テストデータでは評価値 1 がつけられている評価値は予測後の評価行列では多く「使いにくい」から「どちらかといえば使いにくい」の評価と推定される．また，評価値 4 が「使いにくい」の評価と推定される例が少ない．これにより，重み付き分解 A では「使いにくい」と評価された Web ページが「どちらかといえば使いやすい」「使いやすい」に推定される問題が少ないことと，「使いやすい」と評価されている Web ページが「使いにくい」と推定される問題も同時に少ないことが確認された．

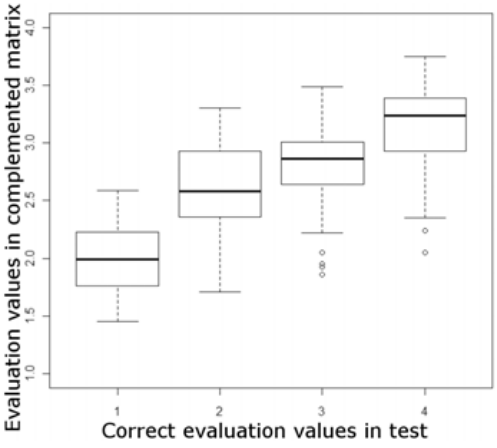


図 1 テストデータの実評価値と予測後の推定評価値（予測後の評価値）の関係

Fig. 1 Relation of the evaluation values with practical testing and complementation for testing data.

5. 考察

本章では行列因子分解による Web ユーザビリティの評価値行列の予測について考察を述べる．

5.1 評価値の予測の有効性

本節ではユーザビリティテスティングにおいてタスク達成のために閲覧する Web ページが被験者ごとにどの程度異なるか示し，評価値の予測の有効性を検討する．

本論文で対象にした 8 社のうち，ある企業の評価値行列について考える．この企業の Web ページを閲覧するタスクを行った被験者は 8 名であり，タスク達成までの平均閲覧 Web ページ数は 7.75 PV であり，最短で 5 PV，最長は 11 PV の閲覧でタスクを達成している．また，被験者が 1 PV でも評価したページは 12 種類のページである．それぞれのユーザが閲覧したページは**表 6** のようになる．表 6 において，それぞれページを数字で表しており，同一の数字は同じページを閲覧していることを意味する．また，各列は被験者（A から H）を表し，各被験者が閲覧した順にページを表している．Top ページは “1” であり，目的の情報が記されたページは “5” であり，これにたどり着くことでタスク達成としている．

ここで，ページ “8”，“9”，“10”，“11”，“12” はそれぞれ 1 人のユーザしか閲覧していない．このようなページは評価値の信憑性が低いため，分析対象ページから外す，またはタスク実行数を増やし多くの被験者に閲覧させるという対処法が考えられ，ユーザテストのコストが無駄になる場合がある．一方で，これらのページに対するユーザビリティ評価値は利用されない場合，ユーザビリティに関する問題点が見落とされる可能性がある．たとえばページ “8” において被験者 C は「使いにくい」と評価しており「ページが変化した（遷移した）ことに気づかなかった」「目的のリンク表示が小さく気づかなかった」といったユーザビリティに関する問題点についての意見を述べている．もし被験者 C と似た評価傾向を持つ被験者であれば，同様にページ “8” に遷移する可能性もあり，こうした少数の被験者しか評価していない Web ページに対してユーザビリティ評価を実施できる可能性も高まる．

表 6 被験者ごとの Web ページ閲覧順序

Table 6 Browsing order of Web pages for every subject.

閲覧順	被験者							
	A	B	C	D	E	F	G	H
1PV 目	1	1	1	1	1	1	1	1
2 PV 目	2	2	2	2	11	2	2	12
3 PV 目	3	6	6	6	2	6	6	2
4 PV 目	4	3	8	3	6	3	2	6
5 PV 目	5	4	3	4	3	4	3	3
6 PV 目		7	4	5	4	7	4	4
7 PV 目		5	9		4	4	5	7
8PV 目			10		7	5		5
9PV 目			4		5			
10PV 目			7					
11PV 目			5					

事前に 1 つ以上のタスクが終了していれば，それらの評価値と他の被験者が終了したタスクの評価値を用いて行列因子分解による評価値行列の予測を行うことにより，実施していないタスクにおける評価値を推定することが可能である．これにより未実施のタスクの中からその被験者がより多くのユーザビリティ問題を抽出できることが期待されるタスクを推薦することが可能になる．

5.2 予測前後の評価値の相関

予測した精度はテストデータとの比較により測ることができ，重み付き分解 A による分解の精度が良い結果となった．一方で予測前の限られた被験者数による評価値行列と予測後のすべての被験者による評価値行列を比較した場合，予測前後で評価傾向が変わらない必要がある．

本節では，予測後の評価値行列は，未知の評価値も含めて，予測前の評価値と評価傾向が変わらないことを確認する．そのため，重み付き分解 A を用いて，あるデータセットに対し評価値行列の予測を行った場合において，予測前の評価値行列の各 Web ページの平均評価値を比較する．それぞれの Web ページの平均評価値のヒストグラムは**図 2**，**図 3** のようになる．

予測後の評価値行列が予測前の全評価値の平均値である $r = 2.90$ 付近に多く集中しているように見られる．しかし，予測前と予測後の Web ページの平均評価値の相関は 0.826 であり，無相関を帰無仮説とした相関の検定を行ったところ P 値は 2.2×10^{-16} 以下であり，強い相関がみられる．以上より予測前の欠損値を多く含む評価値行列と予測後の評価値行列で各 Web ページに対する評価傾向は変

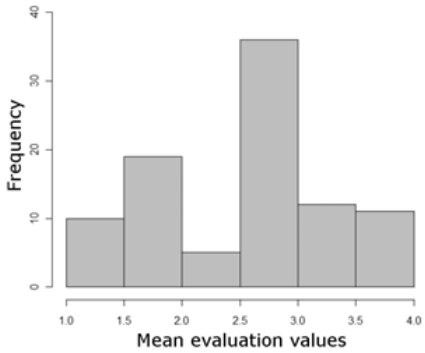


図 2 予測前のページ別平均評価値のヒストグラム
Fig. 2 Histogram of the mean evaluation values for every Web pages before complementation.

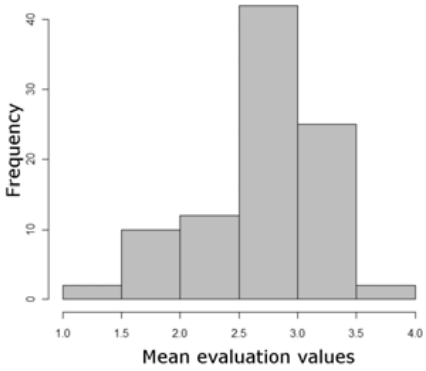


図 3 予測後のページ別平均評価値のヒストグラム
Fig. 3 Histogram of the mean evaluation values for every Web pages after complementation.

わらないことが分かった．

5.3 予測後の評価値行列を用いた評価

予測後の評価値行列は，予測前の評価値行列では分析できなかったユーザビリティ評価に関する分析が可能になることが望まれる．本節では，ある Web ページに対し予測前の限られた被験者数による評価値と，予測後のすべての被験者による評価値を比較し，予測後の評価値行列を用いたユーザビリティ評価について検討する．

本節で扱う Web ページは 5 名の被験者が評価しており，2 名が評価 1，2 名が評価 2，1 名が評価 4 をつけているページがある．このページにおけるすべての被験者の予測後の評価値を見ると**図 4** のようになる．

図 4 において，縦軸は予測後の評価値を表し，横軸は 35 名の被験者を表す．図中の黒棒で表された棒グラフは実際にこの Web ページを評価した被験者の実際の評価値である．予測後の評価値において四捨五入を行い 4 段階評価に変換すると，評価 1 をつける被験者は 4 名，評価 2 が 14 名，評価 3 が 14 名，評価 4 が 3 名であり，予測前に評価 1 をつけていた被験者のほかに 1 名がこの Web ページに「使いにくい」とつける可能性があると推定され，6 名が「どちらかといえば使いにくい」とつける可能性がある

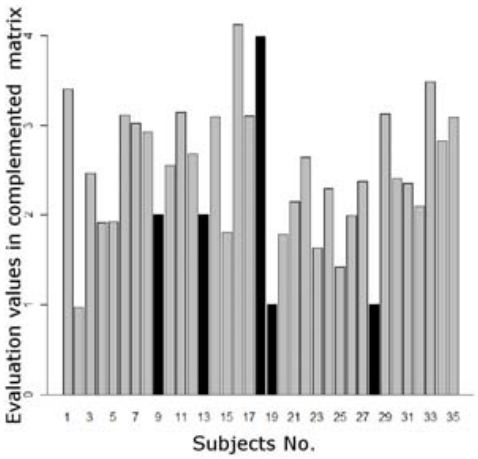


図 4 ある Web ページに対するすべての被験者による評価値
Fig. 4 Evaluation values for a Web page with all subject.

ことが分かる．以上のように，予測前では一部の被験者の突出して低い評価値であっても，予測後の評価値行列において，被験者の予測評価値と比較することで，同様に低い評価値をつける被験者を発見することも可能であり，ユーザビリティ問題個所の抽出のために低い評価をするであろう Web ページを多く含むタスクを実施するよう調整することも可能である．このように限られた被験者数でも，多くのユーザビリティの問題点を発見できるようなユーザビリティテスティングを実施し，多くの被験者からの評価を得ると同様の分析が可能となると考える．一方で評価値そのものは推定評価値であり，実際に把握した主観的満足度の評定値そのものではない．そのため，本手法による数値を用いて追加的なタスクを提案し，実際に被験者の評価を得ることが重要であると考える．

6. まとめ

Web サイトを対象とするユーザビリティテスティングにおいて，少数の被験者しか評価していない Web ページはユーザビリティ評価の分析対象とするか判断が難しい．本論文では，他の被験者が訪れた場合の評価値を判断材料として加えるため，協調フィルタリングの手法の 1 つである行列因子分解を用いて，ユーザビリティの評価値行列の欠損値の予測を行った．

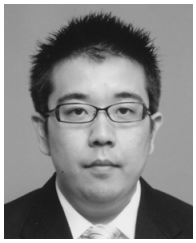
実際にユーザビリティテスティングを実施し，収集した Web ユーザビリティの評価値行列は 84%欠損値を含んでいた．この評価値行列に対し，基本的な行列因子分解に加え，バイアス付き行列因子分解と，評価値の重み付き行列因子分解を用いた予測を試みた．行列因子分解法による予測実験を行った結果，平均評価値から離れた評価値を重視するメリット関数を用いた行列因子分解による予測法が最も優れた精度で予測することができた．またユーザビリティの 4 段階評価において 1 段階以下の誤差で評価値の予測が可能であった．今後，評価の低いページの予測精度を

向上させることは課題として残されている．これは行列因子分解における重み付き分解での重みパラメータのとり方を，今回は全体の予測精度を基に選択したが，評価の低い Web ページの予測値の精度を基に選択することや，評価の低い Web ページの影響をより強くするように重みを与えることで改善されるものと考える．

Web サイトに対するユーザビリティテスティングにおいて，被験者が評価完了した一部の Web ページの評価を用いて重み付き行列因子分解により予測した被験者の評価傾向から，未評価の Web ページであっても低い評価をする Web ページを推定することが可能である．ユーザビリティテストを実施している現場において，そのような情報を得ることができれば，追加的に実施するタスクを選定し，被験者に実際に評価を行ってもらうことによって，Web サイトのユーザビリティ上の問題点のあるページを，より多くかつ的確に把握するよう支援することが可能となる．

参考文献

[1] Amazon.com, available from ⟨http://www.amazon.com/⟩.
[2] Dumas, J.S. and Redish, J.C.: *A Practical Guide to Usability Testing*, Ablex Publishing (1993).
[3] Goldberg, D., Nichols, D., Oki, B.M. and Terry, D.: Using Collaborative Filtering to Weave an Information Tapestry, *Comm. ACM*, Vol.35, pp.61–70 (1992).
[4] Herlocker, J., Koren, J. and Riedl, J.: Explaining Collaborative Filtering Recommendations, *ACM, 2000 Conference on Computer Supported Cooperative Work*, pp.241–250 (2000).
[5] Hu, Y., Konstan, Y. and Volinsky, C.: Collaborative Filtering for Implicit Feedback Datasets, *Proc. IEEE Int'l Conf. Data Mining* (*ICDM '08*), pp.263–272, IEEE CS Press (2008).
[6] 北島宗雄：ユーザビリティテスティングについて，情報の科学と技術，情報科学技術協会，Vol.54, No.8, pp.391–397 (2004).
[7] Koren, Y., Bell, R. and Volinsky, C.: Matrix Factorization Techniques for Recommender Systems, *IEEE Computer*, Vol.42, No.8, pp.30–37 (2009).
[8] 中道 上，木浦幹雄，山田俊哉，上野秀剛：Web インタラクションの協調的可視化ツールの提案，ヒューマンインタフェースシンポジウム 2010 論文集（DVD-ROM），pp.341–344 (2010).
[9] Netflix: Netflix – Watch TV Shows Online, Watch Movies Online, available from ⟨http://www.netflix.com/⟩.
[10] Netflix Prize: Netflix prize Rules, available from ⟨http://www.netflixprize.com//rules⟩.
[11] Nielsen, J.: *Designing Web Usability*, Peachpit Press (1999).
[12] Nielsen, J.: *Usability Engineering*, Academic Press London (1993).
[13] Nielsen, J.: Iterative User-Interface Design, *IEEE Computer*, Vol.26, No.11, pp.32–41 (1993).
[14] Nielsen, J. and Pernice, K.: *Eye tracking Web Usability*, New Riders Press (2009).
[15] 日本ブランド戦略研究所：Web サイト評価ランキング 2010，入手先 ⟨http://japanbrand.jp/ranking/we-ranking/we2010.html⟩.
[16] Paterek, A.: Improving regularized Singular Value Decomposition for Collaborative Filterings, *Proc. KDD Cup and Workshop*, pp.39–42, ACM Press (2007).
[17] Resnick, P., Lacovou, N., Suchak, M., Bergstrom, P. and Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews, *Proc. CSCW '94*, pp.175–186 (1994).
[18] Shardanand, U. and Maes, P.: Social information filtering: algorithms for automating "word of mouth", *Proc. CHI '95*, pp.210–217 (1995).



山田 俊哉

2009 年南山大学数理情報学研究科博士前期課程修了．2013 年総合研究大学院大学複合科学研究科統計科学専攻博士課程修了．同年 NTT アイティ株式会社入社．博士（学術）．計算機統計学会会員．

中道 上 （正会員）

2004 年奈良先端科学技術大学院大学博士前期課程修了，2007 年同大学博士後期課程修了，1999 年住友金属システム開発株式会社（現，アイエス情報システム株式会社）入社，2007 年南山大学数理情報学部講師，2013 年福山大学工学部情報工学科准教授，博士（工学）．ソフトウェアユーザビリティ評価，インタラクション分析の研究に従事．電子情報通信学会，IEEE，ヒューマンインタフェース学会各会員．

松井 知子

1988 年東京工業大学大学院修士課程修了．同年 NTT（株）入社．話者・音声認識の研究に従事．1998 年より ATR 音声翻訳通信研究所，2000 年より ATR 音声言語通信研究所および音声言語コミュニケーション研究所に出向．2001 年 1〜6 月米ルーセント・テクノロジー社ベル研究所客員研究員．2003 年より情報・システム研究機構統計数理研究所准教授．2008 年より同研究所教授．統計数理の研究に従事．東京工業大学博士（工学）．IEEE，日本音響学会，日本統計学会各会員．1993 年電子情報通信学会論文賞受賞．