Vol.2013-DPS-156 No.24
Vol.2013-GN-89 No.24
Vol.2013-EIP-61 No.24
2013/9/13

行列因子分解による Web ページのユーザビリティ評価値の補完

山田俊哉[†1]　中道上[†2]　松井知子[†3]

Web ページのユーザビリティテスティングにおいて，限られた人数の被験者となるユーザによるユーザビリティについての評価を有効に利用することを目的として，行列因子分解を用いて被験者が未評価の Web ページの評価値を補完する方法を提案する．Web ページのユーザビリティテスティングでは，閲覧する Web ページの順序や種類が被験者により異なる場合がある．そのため評価対象となる全ての Web ページに対し，全ての被験者の評価値を得ることが難しい．行列因子分解は（Web ページ数）×（ユーザ数）で表わ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り，本手法により評価値行列の欠損値を補完することが可能となる．基本的な行列因子分子に加え，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補完法を試みた．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補完法を適用した結果，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補完が可能であった．

Score Complementation in Web Usability Evaluation Using Matrix Factorization

TOSHIYA YAMADA[†1]　NOBORU NAKAMICHI[†2]
TOMOKO MATSUI[†3]

Our goal is to make efficient use of the limited number of the evaluation scores for Web usability by pre-determinate subjects. Currently there are an enormous number of Web sites and each site includes multiple pages. The method to evaluate Web usability by subjects and analyze the problems in each page is demanded. Since it is difficult for each subject to evaluate a large number of all pages in the target sites, some evaluation scores are missed. In this paper, we propose a method to complement the missing scores based on the matrix factorization technique. It is experimentally shown that our method obtains the complementation error which is less than one in assessment in four grads.

1. はじめに

近年，Web 上では様々な情報が発信され，企業においても Web サイトの事業貢献度は高まっている．その中で Web サイトの価値も高まり，数百億円を超える経済価値を持つ Web サイトも存在する[14]．そのため，Web ページの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，ユーザビリティに関する問題を見つけるユーザビリティテスティングが行われている[11][12].

Web ページ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が，実施するにあたって被験者を集めるコストが大きい問題が挙げられる[2]．実際に被験者が使用するため評価対象となる 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 Web サイトで特定の情報を発見するなどのタスクを課す[6]．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，種類は被験者毎に異なるため，一人の被験者が全ての評価対象となる Web ページを評価することは難しい．そのため，（Web ページ数）×（ユーザ数）の評価値行列に欠損値を多く含む問題がある．そのため，その Web ページの評価値を無駄にしてしまうといった問題や，ユーザビリティテスティングに新たな被験者を加えなければならない問題が生じる．

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[16][17].

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

†1 NTT アイティ株式会社
NTT IT CORPORATION
†2 福山大学
Fukuyama University
†3 情報・システム研究機構 数理研究所
The Institute of Statistical Mathematics

Vol.2013-DPS-156 No.24
Vol.2013-GN-89 No.24
Vol.2013-EIP-61 No.24
2013/9/13

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ディション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，ユーザビリティテスティングにおいて，これまで分析対象として用いられていなかった少数の被験者しか評価していない Web ページもユーザビリティの分析対象として利用することを目的とする．そのため，欠損値を補完し，ある被験者が未閲覧の Web ページの評価値を推定するとともに，評価値行列からユーザまたは Web ページに関する潜在因子を抽出し，ユーザビリティの分析に役立つ可能性がある情報を提供する方法を提案する．具体的には，欠損値を多く含む（Web ページ数）×（ユーザ数）の評価値行列に対し，協調フィルタリング手法の一つである行列因子分解を用い，被験者が未評価の Web ページの評価値を推定し，限られた被験者数によるユーザビリティテスティングの評価値を有効に利用することを試みる．

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補完法について述べる．4 章では行列因子分解による評価値行列の補完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][13].

Jakob Nielsen は，ユーザビリティの重要な要因の一つとして，ユーザによる主観的満足度を挙げている[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様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ためのユーザビリティテスティングを想定し，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，評価値を得るため，被験者が閲覧した Web ページの数を PV（Page View）としてカウントする．

3. 行列因子分解による補完法

協調フィルタリング手法の一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補完を行う．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補完法について述べる．

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．

$$r_{ui} = q_i^T p_u \qquad (1)$$

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．

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

κ は入力データ中に存在する(u, i)の全てのペアであり，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良く知られており，これをバイアス項として行列因子分解に用いる[15]．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．

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

ここで μ は入力データにおける欠損値ではない全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．

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

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

Vol.2013-DPS-156 No.24
Vol.2013-GN-89 No.24
Vol.2013-EIP-61 No.24
2013/9/13

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補完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．

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

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よ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 つの種類で表わし，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

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

重み B: 被験者毎の評価値の平均値 μ_u からの距離

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

重み C: Web ページ毎の評価値の平均値 μ_i からの距離

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

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」，被験者毎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 ページ毎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. 評価値行列の補完実験

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補完実験について述べる．はじめに，補完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，被験者及び評価対象の 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補完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

補完実験には，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通りである．

- ディスプレイ: 液晶 21 インチ(有効表示領域: 縦 30cm, 横 40cm，解像度: 1280pixel×1024pixel)
- 顔とディスプレイの距離: 約 50cm
- 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 つの企業サイトの内，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．一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 1PV 毎に被験者のユーザビリティ評価値を得た．

Vol.2013-DPS-156 No.24
Vol.2013-GN-89 No.24
Vol.2013-EIP-61 No.24
2013/9/13

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 1PV 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．たとえば被験者 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 とする．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ユーザが付ける評価値を１PV 分に集約する必要がある．そのため一人の被験者が同じ 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，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 498PV であり，そのうち「使いにくい」=59PV，「どちらかといえば使いにくい」=113PV，「どちらかといえば使いやすい」=145PV，「使いやすい」=181PV である．

被験者 35 名，93 ページにおいて 498PV 分の評価値であるため，93 行×35 列の評価値行列の内，498 箇所に評価値が存在し，残りの 2757 箇所が欠損値となる．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補完結果

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 つの補完法と，特異値分解(SVD)および GroupLens 法により補完を試みた．

まずそれぞれの補完法において，パラメータを決定するため予備実験を行う．予備実験では補完法のパラメータであるλを 0.001≦λ≦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補完した結果を挙げる．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補完後の評価値行列の一部であり，欠損値の無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λ=0.1 とした．

表 1　補完前の評価値行列（一部），“-”は欠損値
Table 1　A part of the evaluation matrix before complementation. (“-”; 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　補完後の評価値行列(一部)，**太字**は補完した値
Table 2　A part of the evaluation matrix before complementation. (**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

次に補完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及び GroupLens 法と行列因子分解を用いた補完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補完し，その補完した評価値行列とテストデータの評価値を比較し補完精度を測った．

学習データには評価値行列の中の欠損値ではない評価

Vol.2013-DPS-156 No.24
Vol.2013-GN-89 No.24
Vol.2013-EIP-61 No.24
2013/9/13

値の 2/3 にあたる 332PV 分の評価値をランダムに抽出したものを用い，残りの 166PV 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 74PV 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補完を行い，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補完後の評価値の平均二乗誤差であり，式(9)で定義する．

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

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 を用いて各補完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: λ=0.4，バイアス付き分解: λ=0.4，重み付き分解 A=重み付き分解 B=重み付き分解 C: λ=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．

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

補完法	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

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,…,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 の補完精度が最も高いことがわかった．

一方で，補完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補完した評価値行列の平均評価値は表 4 のようになる．

表 4　100 組のデータセットの平均 RMSE
Table 4　Mean RMSE values for 100 date sets.

補完法	平均評価値(補完前)	平均評価値(補完後)
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

それぞれの手法における補完後の平均評価値の差を t 検定で確認すると，全ての手法の組み合わせにおいて有意水準 5%で差が確認された．表 4 より，補完前と補完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補完する傾向がある．また重み付き分解 A は RMSE が最も小さく補完精度は高いが, 補完前に比べて平均評価値は基本分解に次いで補完前と補完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（補完した評価値）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 5　100 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.

補完法	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 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わかる．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 などに比べて補完の精度は低い．また，重み付き

Vol.2013-DPS-156 No.24
Vol.2013-GN-89 No.24
Vol.2013-EIP-61 No.24
2013/9/13

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，補完行列で平均 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 を適用した例を挙げる．あるデータセットにおいて，テストデータの実評価値と補完後の推定評価値（補完後の評価値）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，縦軸は補完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 が付けられている評価値は補完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．

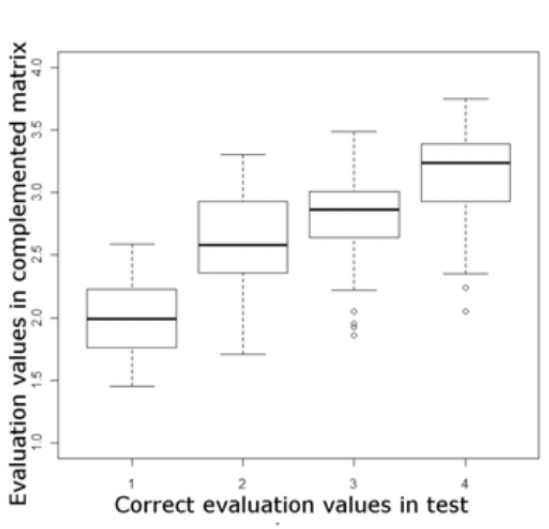

図 1　テストデータの実評価値と補完後の推定評価値（補完後の評価値）の関係

Figure 1　Relation of the evaluation values with practical testing and complementation for testing date.

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補完について考察を述べる．

5.1 評価値の補完の有効性

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 ページが被験者毎にどの程度異なるか示し，評価値の補完の有効性を検討する．

本論文で対象にした 8 社の内，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 7.75PV であり，最短で 5PV，最長は 11PV の閲覧でタスクを達成している．また，被験者が 1PV 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．

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					

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．補完を行わない場合，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る．しかし，分析から外した場合，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ず，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．例えばページ“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ため，このページを分析対象から外してしまうとユーザビリティに関する問題点を見落とす可能性が高まる．一方，タスク実行数を増やす場合，この企業では被験者が平均 7.75PV を閲覧していることから，被験者が分析対象となる 12 種類のページを全て閲覧する可能性は少ない．またユーザビリティテスティングはタスクを遂行する中でのユーザビリティを評価する手法であるため，評価対象ページを全て被験者に閲覧するように指示することは困難である．さらに，新規の被験者にユーザビリティテスティングを実施する場合，分析対象ページが増える可能性もあり，欠損値を減らすことは難しい．

行列因子分解による評価値行列の補完を行うことによ

Vol.2013-DPS-156 No.24
Vol.2013-GN-89 No.24
Vol.2013-EIP-61 No.24
2013/9/13

り，ページ“8”のようなページにも，ユーザビリティテスティングに参加した他の被験者による評価値を推定することが可能である．

5.2 補完前後の評価値の相関

補完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．一方で補完前の限られた被験者数による評価値行列と補完後の全ての被験者による評価値行列を比較した場合，補完前後で評価傾向が変わらない必要がある．

本節では，補完後の評価値行列は，未知の評価値も含めて，補完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補完を行った場合において，補完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．

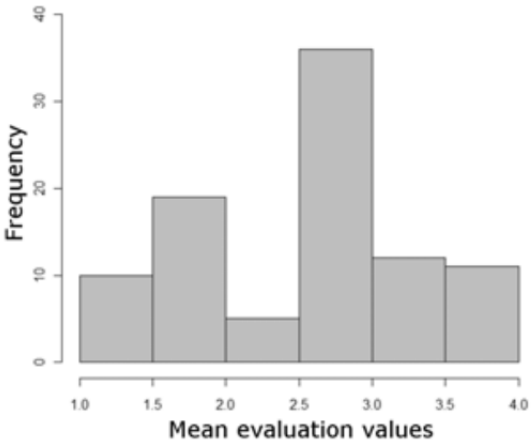

図 2　補完前のページ別平均評価値のヒストグラム
Figure 2　Histogram of the mean evaluation values for every Web pages before complementation.

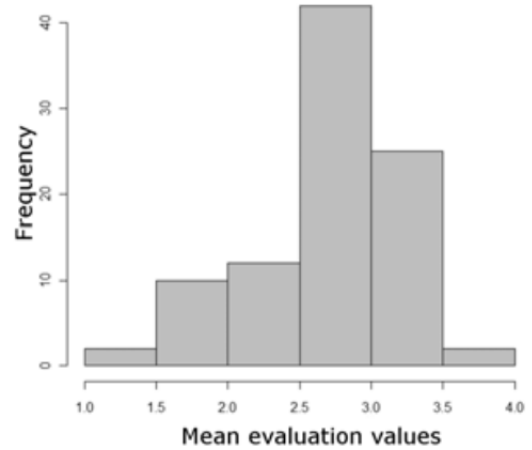

図 3　補完後のページ別平均評価値のヒストグラム
Figure 3　Histogram of the mean evaluation values for every Web pages after complementation.

補完後の評価値行列が補完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，補完前と補完後の 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．以上より補完前の欠損値を多く含む評価値行列と補完後の評価値行列で各 Web ページに対する評価傾向は変わらないことがわかった．

5.3 補完後の評価値行列を用いた評価

補完後の評価値行列は，補完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 ページに対し補完前の限られた被験者数による評価値と，補完後の全ての被験者による評価値を比較し，補完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 1，3 名が評価 2，1 名が評価 4 を付けているページがある．補完後の評価値行列において，全ての被験者によるこのページの評価値を見ると図 4 のようになる．

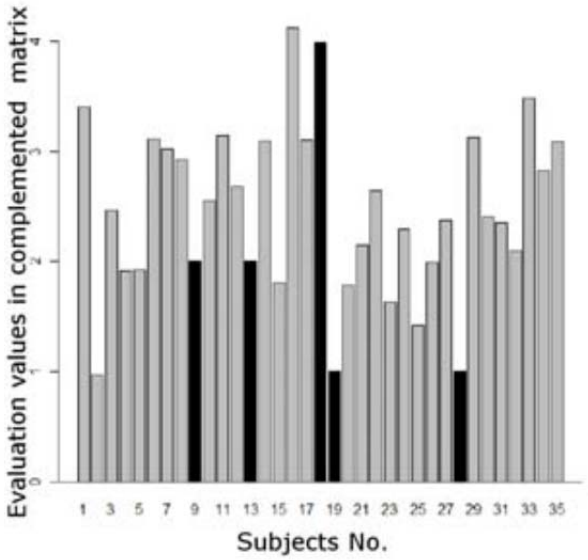

図 4　ある Web ページに対する全ての被験者による評価値
Figure 4　Evaluation values for a Web page with all subject.

図 4 において，縦軸は補完後の評価値を表し，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評価値である．補完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 を付ける被験者は 4 名，評価 2 が 14 名，評価 3 が 14 名，評価 4 が 3 名であり，補完前に評価 4 を付けていた被験者の他に 2 名もこの Web ページに「使いやすい」と付けると推定される．また，評価値の平均も補完前は 2 だったのに対し，2.45 となっており，補完前よりも高くなっ

Vol.2013-DPS-156 No.24
Vol.2013-GN-89 No.24
Vol.2013-EIP-61 No.24
2013/9/13

ているものの，このページは「どちらかといえば使いにくい」ページに入ると考えることができる．以上のように，補完前では一部の被験者の突出した評価値であっても，補完後の評価値行列において，全ての被験者の評価値と比較することで，分析に利用可能である．多くの欠損値を含んでしまう評価値行列を，行列因子分解により補完することで，限られた被験者数でも，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．

まとめ

ユーザビリティテスティングにおいて，限られた被験者数で得られた評価値を有効に利用するため，協調フィルタリングの手法の一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補完を行った．本研究で用いたユーザビリティ評価値データにおいて，行列因子分解法による補完実験を行った結果，平均評価値から離れた評価値を重視して評価値の推定を行うメリット関数を用いた行列因子分解による補完法が最も少ないRMSEで補完することができた．

本手法を用いることで，被験者にタスクの実行を依頼するユーザビリティテスティングによる評価値行列の欠損値を，既知の評価値を用いて補完し，一部の被験者しか閲覧していない Web ページにおいても，未閲覧の被験者の評価値を推定できるため，限られた被験者数によるユーザビリティテスティングの評価値を有効に利用することが可能となる．

参考文献

1) Amazon.com, http://www.amazon.com/
2) J. S. Dumas, J. C. Redish: A Practical Guide to Usability Testing, Ablex Publishing (1993).
3) D. Goldberg, D. Nichols, B. M. Oki, D. Terry, Using Collaborative Filtering toWeave an Information Tapestry, Communicatios of the ACM, Vol.35, pp. 61-70 (1992).
4) J. Herlocker, J. Koren, J. Riedl: Explaining Collaborative Filtering Recommendetions, ACM, 2000 Conference on Computer Supported Cooperative Work, pp. 241-250 (2000).
5) Y. Hu, Y. Konstan, C. Volinsky: Collaborative Filtering for lmplicit Feedback Datasets, Proc. IEEE lnt'l Conf. Data Miing(ICDM 08), IEEE CS Press pp. 263-272 (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) Y. Koren, R. Bell, C. Volinsky: 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, http://www.netflix.com/
10) Netflix Prize: Netflix prize Rules, http://www.netflixprize.com//rules
11) J. Nielsen: Designing Web Usability, Peachpit Press (1999).
12) J. Nielsen: Usability Engineering, Academic Press London (1993).
13) J. Nielsen, K. Pernice: Eyetracking Web Usability, New Riders Press (2009).
14) 日本ブランド戦略研究所: Web サイト評価ランキング 2010, http://japanbrand.jp/ranking/we-ranking/we2010.html
15) A. Paterek: Improving regularized Singular Value Decomposition for Collaborative Filterings, Proc.KDD Cup and Workshop, ACM Press, pp. 39-42 (2007).
16) P. Resnick, N. lacovou, M. Suchak, P. Bergstrom, J. Riedl: GroupLens: an open architecture for collaborative filtering of netnews, CSCW '94 Proceedings of the 1994, pp. 175-186 (1994).
17) U. Shardanand, P. Maes: Social information filtering: algorithms for automating "word of mouth"., CHI '95 Proceedings, pp. 210-217 (1995).