オンライン活動とオフライン活動の相関関係分析

赤池 勇磨[†1,a)]　荒川 豊[†1,b)]　安本 慶一[†1,c)]

概要：近年, 健康支援を目的としたウェアラブルデバイスの利用が活発化している. 本研究の目的は, こうしたセンシングデバイスの利用において, 装着のし忘れや電池切れによって欠損したデータを推定することである. その際, 出来る限りユーザに追加作業による負荷を与えずに, 欠損部分の補完を行うための情報の抽出が求められる. 本研究では運動 (オフライン) とネット閲覧等の (オンライン) 活動に相関関係があると考え, それぞれの活動履歴の抽出, 可視化を行うシステムを構築し, オンラインログがオフラインログの欠損部分の推定に用いられるかについて考察を行う. オンラインログはユーザの活動により自動的に生成されるものであるため, 本手法によるユーザ負担の増加はない. 実際に構築したシステムを用いて, 自動的に被験者 2 名から複数のオンラインログ (LINE, Twitter の利用履歴) とオフラインログ (Fitbit Zip によって取得された歩数データ) を収集し, どういった相関関係や特徴が存在するかを分析した. 結果, オンラインログとオフラインログの間には様々な視点から逆相関関係が見られ, 双方のデータを組み合わせて用いることで, 行動推定がある程度実現可能であることが分かった.

1. はじめに

近年, 肥満, 高血圧, 糖尿病等の生活習慣病患者の増加に伴い, 運動不足や生活習慣の見直しが重要とされている [1][2]. 特に日本では少子高齢化が著しく, 国民全体の健康管理が課題として挙げられる. そうした問題に対して, 医療分野だけに留まらず, 様々な研究やサービスの開発を通した健康改善が古くから試みられている. 健康支援の一環として, 人の運動履歴をセンシングし, 身体情報を推定する健康データマイニングが, 様々な研究分野で活発に行われている [3][4]. 健康データマイニングは, ユーザが身体にコンピュータを身につけることで, ユーザによる行動の告知が無くとも, 様々な行動データを取得する. この行動データから, ユーザがどういった行動を行っていたか, どういう状態に至っていたか等のユーザコンテキスト推定を行うことができる. こうした背景の中, 健康データマイニングを用いた健康支援の為の研究やサービスが普及の兆しを見せている. それらの研究やサービスでは, デバイスを装着することで, ユーザがどういう行動を取ったかを推定し, その運動量等からフィードバックをユーザに返すことを目指している. 健康データマイニングを精密に行うためには, ユーザが常にデバイスを装着し, 運動データが蓄積されることが必須である. しかし, 日々の装着の手間や, 外出前の装着のし忘れ, 電池切れ等が原因で蓄積データに欠損部分が生じるケースは十分に考えられる. また, 蓄積データの欠損はデータマイニングの精度に大きく影響することから, 運動データを用いた診断システムや推薦システムの機能が低下し, それに伴いユーザの継続的な記録に対するモチベーションの低下も起こり得る. よって健康支援に向けたデータマイニングでは, いつも装着されていることが前提であるデバイスが, 電池切れや装着のし忘れによってデータが取れていない時に, その行動ログをいかに推定するかという課題が存在する.

蓄積されるデータの欠損部分を推定し, 適切な値で補う欠損値補完として, 過去データの統計値より得られた値による補完手法がいくつか提案されている [5][6][7]. しかしながら, 健康支援に向けたデータマイニングでは, その日の行動がその日のデータマイニングに影響を及ぼす. よって, 精度の高いフィードバックをユーザに返すためには, 単なる過去データの平均等ではなく, その日のコンテキスト情報を考慮した過去データによる推定が求められる.

本研究では, ネットサーフィンや SNS への投稿といった人のインターネット上での活動 (以降オンライン活動と定義) と, 歩行等の現実世界での活動 (以降オフライン活動と定義) に何らかの相関関係があると考え, それらを収集, 統合, 分析を行うシステムを構築し, 比較を行う.

オンライン活動とは, Web 上のサービスの利用内容やそ

[†1] 現在,
Presently with 奈良先端科学技術大学院大学 NAra Institute of Science and Technology
a) akaike.yuma.aj6@is.naist.jp
b) ara@is.naist.jp
c) yasumoto@is.naist.jp

表 1　一般的なオンラインログと考えられるユーザの行動情報

	Gmail	LINE	Twitter	Facebook(メッセージ投稿)	Facebook(チェックイン)	Foursquare	Blog	Chat	Evernote
在室中	◎	○	△	◎	△	△	◎	◎	◎
外出中	×	×	○	△	◎	◎	×	×	○

の頻度のことを指し, WebMail サービスや, ソーシャルネットワークサービス (SNS) 等の利用履歴がそれに該当する. 以下, オンライン活動による履歴をオンラインログ, オフライン活動による履歴をオフラインログと呼ぶ. こういった, 生活中に自然に蓄積される情報はライフログと呼ばれることが多く, ユーザに負担をかけない暗黙的プロファイル生成の方法として, 近年注目されている [8]. オンライン活動とオフライン活動の間に関係性を見つけることができれば, 従来手法と比べてより正確な欠損値の補完を行うことが可能であり, 運動データを用いた診断システム, 推薦システムの精度を向上させることができると考えられる. また将来的には, ウェアラブルデバイスを装着せずとも, 日々の運動量をログできる可能性も考えられる.

本研究では, 研究目的を以下の 3 段階に分けている.

(1) オンラインログ, オフラインログの収集を行うシステムの構築
(2) 本システムより得られたログの可視化, 特徴点の発見
(3) 本システムより得られたログを用いた欠損したデータの推定

本稿では第一段階, 第二段階について取り組み, オンラインログとオフラインログの間に, 将来的に欠損値の補完に用いられる様な関係性が存在するかについて考察を行う.

表 1 は, 一般的に用いられているオンラインサービスの利用から生まれるオンラインログと, ユーザの行動から生まれるオフライン活動ログの予想されうる関係性を示している. オンラインサービスが利用される状況には, ユーザによって偏りがあると考えられ, これらのログを分析することで, 様々な場所におけるユーザの行動を推定できると考える.

オンライン活動, オフライン活動の間の相関関係を求めるため, 被験者 2 人に対し, Fitibit を 1 ヶ月間装着してもらうと共に, オンラインサービスの中でも代表的である Gmail [*1] , Facebook[*2] , Line[*3] , Twitter[*4] におけるオンラインログを提供してもらい, 提案システムを用いて収集, 統合, 分析を行った. 各活動をしたか, していないかで二値化した場合のオンラインログ, オフラインログの相関係数を求めた結果, 使用頻度が最も多かった LINE では-0.79 , 次に使用頻度の多かった Twitter では -0.94 の逆相関が確認でき, 用途の種類にもよると考えられるが, オフライン活動とオンライン活動の関係性が考察できた. これらの結果から, 各ログにて見られる個人の特徴を考慮し, 外れ値を除くことで, 片方のログにおける欠損部分の補完に対して, もう片方のログを利用した推定がある程度可能であることが分かった.

2. 研究背景

2.1 運動 (行動) センシングの動向

これまで 3 軸加速度センサや心拍計測器, 位置情報センサ等を用いた運動センシングに関する研究は広く行われている. 古くは専用のセンシングデバイスを用いた研究が多くなされているが, こうした機材は特殊性, 大きさ, 価格といった面から広く一般的に普及するものではなかった.

しかしながら, 高性能なスマートフォンの普及に伴い, 運動センシングや行動センシングは一般的なものになりつつある. その結果, 一般のユーザ向けのサービスとして, スマートフォンに内蔵されている加速度センサを用いたアプリケーションサービスが普及し始めている [9]. しかしセンシングを行っている間は常にアプリケーションを起動することになるため, こうした携帯型のデバイスでは電力の消費が著しく, 長期に渡るセンシングには向いていない. またランニングやスポーツ時などの激しい運動の計測を行う場合, 複数のセンサーを持ち合わせる研究向けのセンシングデバイスと比べると軽量化されているものの, 装着しながら行動することが難しいケースや, 常に装着できないケースもあり得る. データマイニングの観点から見れば, 蓄積データが揃っていることが前提とされるので, こうした携帯端末によるセンシングは, 健康支援のためのサービスには適していないと言える.

こうした背景のもと, 近年, より小型で長時間駆動可能, そして低価格な運動センシング専用デバイスが登場し普及しつつある. 具体的には, Fitbit Zip[*5], Fuelband[*6], misfit Shine[*7], Jawbone Up[*8] 等が市販されている. この種のデバイスは加速度センサーを搭載して歩数計の役割を持つ他, センシングする対象を最小限にすることで電力の消費を抑えつつ, 消費カロリーの推定や位置情報の記録, 脈拍推定等が可能となっている.

図 1 に示す最も長期的に駆動する Fitbit Zip の場合, 基本的な取得データは加速度センサによる歩数に加え, そこ

*1 www.gmail.com
*2 www.facebook.com
*3 www.line.naver.jp
*4 www.twitter.com
*5 www.fitbit.com.jp
*6 www.nike.com
*7 www.misfitwearables.com
*8 www.jawbone.com

から算出される消費カロリーや階段を登った段数等であり，専用のデバイスと比べると少ない．しかし，スマートフォンや PC に接続することで，測定結果を集計し，図 2 に示す様に分析結果をデバイス，または Web 上で運動頻度を時系列データとして閲覧することができる．また，その軽量さから，常に持ち歩くデバイスとしてユーザへの負担が小さく，消費電力も少ない．こういった点からも，継続的にセンシングを行うことができ，健康支援の為のデータマイニングに優れていると言える．Fitbit Zip ではボタン電池を採用しており，僅か 8 グラムでありながら，一個の電池で 4 〜6 ヶ月もの間，連続したセンシングを行うことができる．本研究では，歩数データの取得の際，この Fitbit Zip をセンシングデバイスとして用いる．

図 1　Fitbit zip

2.2　統計による欠損値補完

しかしながら，最新のウェアブルデバイスを用いても，蓄積データの欠損は発生する．代表的な発生理由として，外出時におけるデバイスの付け忘れが挙げられる．データマイニングにおいては，出来る限り蓄積データは揃っていることが望ましく，揃っていなければ精度の高いデータマイニングは望めない．健康支援に向けられたデータマイニングシステムにおいては，システムが精度の高いフィードバックをユーザに返すことができなければ，ユーザのデバイス装着を継続するモチベーションにさえ影響を与えてしまう可能性があると考えられる．こうしたデータマイニングにおける欠損値の対処として，欠損値補完手法が挙げられる [5]．欠損値補完とは，蓄積データ内で欠落している部分のデータを，一番もっともらしい値で補うことを指す．データマイニングにおける欠損値の補完については，長年研究が行われ，様々な手法が存在する．

簡易な方法として，平均値や中央値といった統計量で補完する方法 (mean imputation) が挙げられる [5]．例として，とある 1 日のデータが欠損していた場合，時系列毎に過去数日間のデータの平均値，中央値を用いて補完を行う．この手法を用いることで，ユーザの生活パターンに合った補完値を推測することができる．しかしこの手法では補完に適さないデータ群も補完に使われてしまう可能性がある．例えば，外れ値や異常値が含まれている場合は，自然な

図 2　fitbit における運動データの可視化

データとは大きくずれた解で補完されてしまう．また，歩数データに対しては問題ないが，YES or NO や男性 or 女性といった 2 値データの補完には用いることはできない．

より真のデータに近い値で補完を行う手法として，よく似た個体の値で補完する方法 (HotDeck Imputation) がある [6]．例として，午後の運動データが欠損していた場合，午前中の運動データの形が似ている 1 日分の午後のデータを補完値として採用する．類似データが必要であるため，センシング初期に用いることは難しいが，この手法を用いることで，Mean Imputation にて問題とされていた，外れ値や異常値を含む補完の緩和を狙うことができる．

他にも，回帰により予測した値で補完する方法 (Regression Imputation) や，母数モデルを仮定して，最尤法あるいはベイズ法を元にした，モデルに基づく方法 (model-based produres) が存在する [7]．しかし，前者は他の変数を用いるので，歩行データのみを取る Fitbit Zip を用いたセンシングには適用が難しく，後者は精度が過去データの量に依存することから，単純に運動データの補完方法として用いることは難しい．

こういった手法を用いることで，データの欠損部分を過去のデータの統計的観点から見た場合に一番真値に近いであろう値で欠損値で補完することが出来る．しかし，健康データマイニングを行う場合，その日の天候やユーザの予定，体調等のコンテキストによって，補完値の求め方は変動すると考えられる．欠損値が発生し補完を行う日が，普段行わないスケジュールを持つ日であった場合は，過去の統計データを用いて欠損値補完を行う際，真の値とはずれた値で補完することになる．例えば，同じ平日の昼間であった

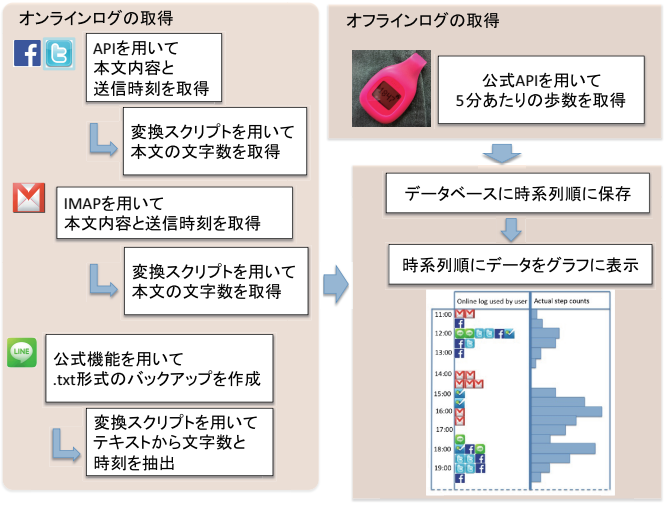

図 3　提案システムの構成

としても, また午前中に同じ様な動きをしていた日があったとしても, 必ずしも同じ時間に食事を摂るというわけではなく, その日のユーザのスケジュールによっては, 食事を摂らずに動き回っている可能性もある. よって, 統計的なデータを用いた欠損値補完では, ユーザのその日の行動に適した補完は難しく, その日そのユーザが行った独特の行動というものを考慮することはできない. こうしたその日のユーザのコンテキストを踏まえた補完を行う為には, 統計的な補完を行う前に, ある程度ユーザが行った行動を推定する必要がある.

また他にも欠損部分を補う方法として, ユーザに直接その日の天候や自身の気分といった, その日独自のコンテキスト情報を入力してもらうよう促す明示的プロファイル生成を用いることで, 精度の高い補完を行うことができると考えられる. しかし, 逐一その日あったことを記載してもらうことは, ユーザの負担となり, センシングという観点からいえば適していないと考えられる.

本研究では, ユーザの負担なく, ユーザの行動と相関を持つ情報をデータ補完に用いられないかと考え, 日々暗黙的に生成されるユーザのオンライン行動履歴 (ログ) に着目する.

3. オンラインログ, オフライン活動の可視化システム

本研究では, センシングができない際のユーザの行動推定にあたり, オフライン活動とオンライン活動の間に何かしらの相関関係があるのではないかと考え, 実際に複数のユーザからオンラインログと歩数データの抽出を計り, 可視化による比較を行えるシステムを構築した. 図 3 は提案システムの構成のイメージ図である.

本研究では, 被験者 2 名 (大学院生) から, 1 ヶ月分の歩数データと, オンラインログを取得し, その相関を確認, 考察する.

提案システムは, 主に 1 ユーザに対する 1 日分のオンラインログ, オフラインログを可視化する. 実装内容について, オンラインログ, オフラインログに分けて説明する.

3.1 オンラインログの抽出

オンラインログとして, 表 1 の中でもより一般的なデータソースである, LINE, Twitter, Facebook, Gmail の以下の 4 種類のオンライン上で得られる情報を用いる.

(1) LINE の送信メッセージ情報 (本文の文字数, 送信時刻)
(2) Twitter のツイート情報 (本文の文字数, ツイート時刻)
(3) Facebook のつぶやき情報 (本文の文字数, 送信時刻)
(4) Gmail の送信メール情報 (本文の文字数, 送信時刻)

LINE では Gmail と異なり, 一般向けの公開データベースやプロトコルは用意されていない. しかし公式に, ユーザ個人用のバックアップ機能が存在し, 本文と送信時間, 宛先を.txt ファイルとして保存することができる. よって期間内に送信されたメッセージをユーザに .txt ファイルに変換してもらい, 提供してもらった. ユーザのプライバシーを考慮し, 文章を文字数へと変換するため, 提案システム側で本文と送信時刻を Python を用いて正規表現し, 本文を文字数へと変換した. その後, 提案システム側で用意したデータベースに保存する.

Twitter におけるユーザのツイート情報は, 公式 API を用いて取得することができる. 公開されている Python ライブラリを用いて, 指定した期間内にユーザから発信されたツイートの本文と時刻を取得した. ユーザのプライバシーを考慮し, 本文を文字数へと変換した後, 提案システム側で用意したデータベースに保存する.

Facebook におけるユーザのツイート情報は, 公式 API を用いて取得することができる. 公開されている PHP ライブラリを用いて, 指定した期間内のユーザから発信されたツイートの本文と時刻を取得した. ユーザのプライバシーを考慮し, 本文を文字数へと変換した後, 提案システム側で用意したデータベースに保存する.

Gmail におけるユーザの送信メール情報は IMAP を用いて取得する. IMAP とは, インターネットやイントラネット上で, 電子メールを保存しているサーバからメールを受信するためのプロトコルであり, 一般的にはメールの管理や共有に用いられることが多い. これを用いて, 期間内におけるユーザの送信メールの本文と送信時刻を全て取得した. ユーザのプライバシーを考慮し, 本文を文字数へと変換した後, 提案システム側で用意したデータベースに保存する. 実装には PHP を用いた.

各オンラインログが取得でき次第, 全てのログに対して提案システム側で用意したデータベースに「どのオンラインサービスであるか」, 「本文の文字数」, 「送信時刻」をまとめ, 時系列順にソートする.

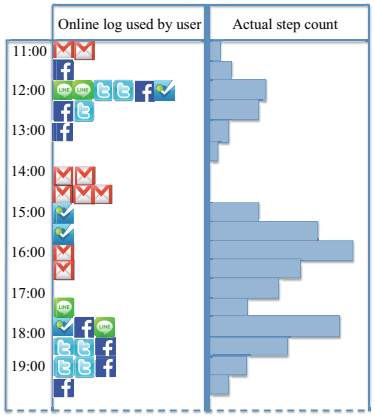

図 4　提案システムの活動量の可視化表のインターフェース

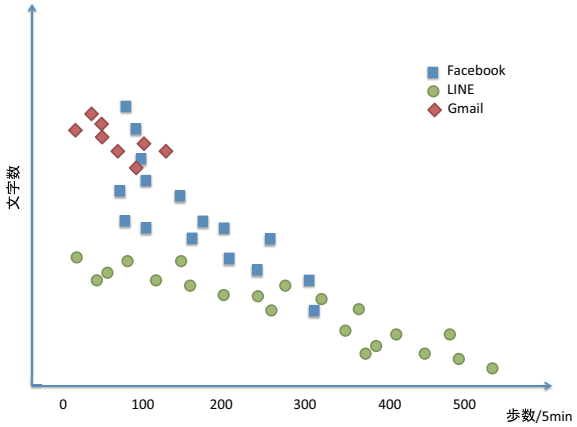

図 5　提案システムのオフライン活動量とオンライン活動量の相関関係を求めるグラフのインタフェース

3.2　オフラインログの抽出

オフラインログとして, fitbit 社の Fitbit Zip にて蓄積された歩数データを用いる. Fitbit Zip では, 単なる万歩計としての歩数データだけでなく, 時間単位の歩数データから推測される動きの活発さや, 消費カロリー等も得ることができる. 本実験では, 5 分あたりの歩数データのみを参照データとして用いる.

Fitbit Zip に蓄積された歩数データは, 専用の bluetooth 受信機をパソコンに装着することで, 簡単にデータをサーバに送信することができる. サーバに保存されているデータは, ユーザアカウント毎に用意されている公式のユーザ専用ページから確認することができる. また, 公開されている FitbitAPI を用いることで, 直接取得することも可能である. 本研究では FitbitAPI を用いて, ユーザに自身のアカウントでログインしてもらうことで, 指定した期間内に蓄積された 5 分あたりの歩数データを取得するプログラムを, 公開 Java ライブラリと併用し, Java にて作成した. 取得された歩数データは, 時系列順にシステム側で用意したデータベースに格納する.

3.3　オンラインログとオフラインログの統合

各ログを抽出し, データベースに格納でき次第, LINE, Twitter, Facebook, Gmail における活動ログを比較するための可視化を行う. Python の matplotlib を用いて, 時系列毎の歩数量のグラフ, 5 分あたりの歩数における文字数量のグラフを作成し, オフラインログのグラフとの比較を行えるよう可視化を行う. 図 4 に示す様に, 前者のグラフでは, 時系列に沿って歩行を行っている時, 行っていない時に, オンライン活動をどの程度行う傾向があるのかを, 直感的に確認することが出来る.

また, 図 5 に示す様に, 後者では 5 分間における活動量が多い時, 少ない時で, オンライン活動との相関関係があるかどうかを見ることができる.

3.4　考察

実際に 2 名の被験者ユーザ (大学院生, 男女 1 名ずつ) から得られた一ヶ月分のデータから, 相関関係の手動での抽出を試み, 本提案システムが実現可能であるかを考察した. 実装したシステムでは, LINE, Twitter, Facebook, Gmail の抽出が可能であるが, Facebook, Gmail については, 1 日あたりで得られるデータ数が極めて少なく, 特徴的な考察が導けなかったので, 本稿では両被験者とも比較的多くのデータが確保できた LINE, Twitter について取り扱った.

まず, オンラインログとオフラインログの間に何かしらの関係性が見られないかを確かめるため, 時系列順に Fitbit Zip より得られたオフラインログと, 各オンラインサービスより得られたオンラインログのデータを並べ, 比較を行った. 図 6 は被験者 A より, 図 7 は被験者 B より得られた 1 ヶ月分のデータの中から, Fitbit Zip, LINE, Twitter の全てにおける活動頻度がある程度多く見られた 1 日を抜き出し, 一般的に活動が盛んであると考えられる午前 10 時から午後 10 時に時間帯を絞り, 数式 (1) に従って正規化した活動量を, 時系列順に並べたものである.

$$Log\ count = \frac{(raw\ data\ of\ Log\ count)}{(Maximum\ raw\ data\ of\ Log\ count)} \quad (1)$$

図 6 から, 被験者 A は午後過ぎから夕方あたりまでのオンライン活動は見られず, 夕方以降に活発なオンライン活動が見られた. この日は平日であったことから, 被験者は午後以降学業もしくは課外活動に務めており, オンライン活動を控えていたと考えられる. また図 7 より, 被験者 B も午後過ぎから夕方にかけてオフライン活動量は少なく, 多くの時間を室内で過ごしていたものと考えられる. また, この日は夕方に LINE から Twitter へのオンライン活動の推移が見られた. またその推移が現れる間には, この日一番多い歩数を示すオフライン活動も観測された. このことから, 大きなオフライン活動により, 生活環境の変化を推測す

ることができると考えられる．具体的には，大学から自宅に帰宅したのではないかと考えられる．このことから，オフラインログ，オンラインログを組み合わせることで，ユーザの行動の中で特徴的な部分を確認することができた．この点から，両ログの関係性は，欠損しているログやその行動量の推定だけに留まらず，1 日のスケジュールの推定にも応用ができると考えられる．また，オフライン活動が活発な時間帯ではオンライン活動はあまり行われず，また逆に，オンライン活動を頻繁に行っている時間帯は，オフラインログによるユーザの移動はあまり確認できないという特徴も見られた．これは，オンライン活動そのものが，移動しながら行うことが難しく，多くの場合停止状態である期間に行われる傾向が強いことを意味しているものと思われる．他にも，オフライン活動がなされた後，オンライン活動が行われる傾向や，決まった時間帯に同じ様なオンライン活動，オフライン活動が行われる等，被験者個人におけるものと考えられる特徴も見られた．これらの結果より客観的観点からも，双方の活動ログの間には相関関係が潜んでいると考えられる．

より具体的に双方の関係性を考察するため，同時間帯に行われたオンライン活動，オフライン活動の相関係数を求めた．オンライン活動及びオフライン活動が行われた時間を 1 ，行われなかった時間を 0 と二値化し，就寝時等の双方の値が 0 になる時間帯を除いた相関係数を求めた．その結果，特に使用頻度の高かった LINE においては，-0.79，次に使用頻度の多かった Twitter では -0.94 と強い逆相関を示しており，この点からも，オンライン活動とオフライン活動には関係性が存在すると考えられる．しかしながら，就寝時間や就業中と考えられる，オンラインログ，オフラインログ共に 0 である時間帯も含めた相関係数は，どのオンライン活動でも 0.1 未満となった．よって，オフラインログと逆相関を持つオンラインログを扱う場合，全時間帯のデータから相関関係を見つけることは難しく，ユーザの活動時間や，周辺の時間帯における活動を考慮し，推定することが必要になると思われる．また，オンラインログの種類によって，オフラインログへの影響力は異なると考えられる．例えば，Twitter や LINE では逆相関が見られたが，位置情報のアップロードを主な目的とした Foursquare[*9] 等におけるオンライン活動では，どちらかと言えば移動中に行われることが多いと思われるため，明確な逆相関関係は見られない可能性がある．さらに，1 日に用いられる各オンライン活動の利用頻度もユーザによって異なると考えられる．このことから，オフラインログの推定や欠損値の補完に応用するにあたっては，個人の利用頻度が及ぼす相関係数値への影響も考慮に入れるべきであると考える．

このように，二値化を行うと相関関係は見やすくなるが，

*9 ja.foursquare.com/

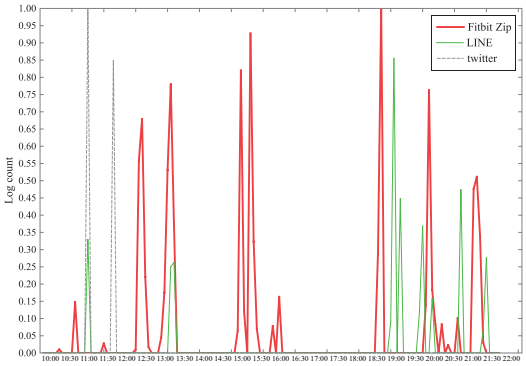

図 6 1 日のオンライン活動量，オフライン活動量の変化 (被験者 A)

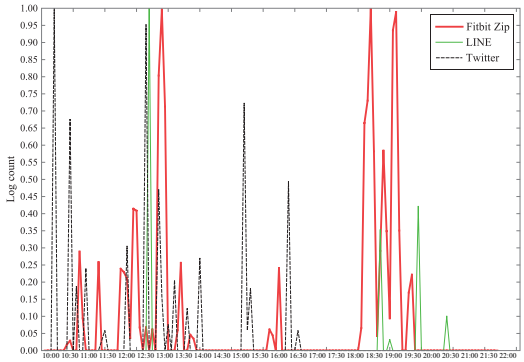

図 7 1 日のオンライン活動量，オフライン活動量の変化 (被験者 B)

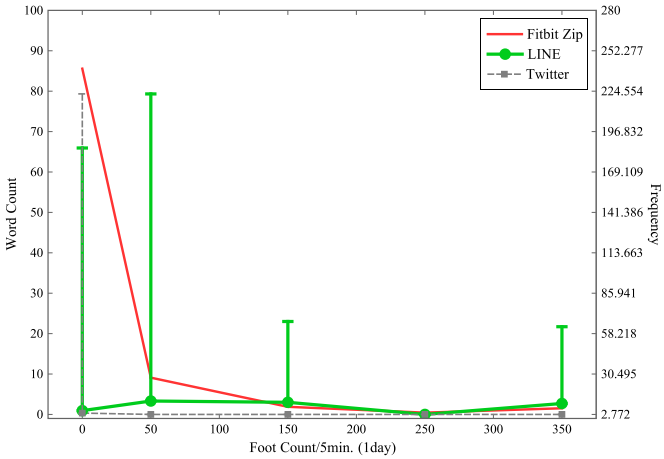

図 8 5 分間におけるオフライン活動量毎の度数分布とオンラインログの頻出頻度 (被験者 A)

ユーザの各活動における特徴点を観察することは難しい．そこで，図 6，図 7 と同日のデータを参照し，各オンラインログの量を，5 分あたりの歩数が 0 歩の時，1〜100 歩の時，101〜200 歩の時，201〜300 歩の時，301〜400 歩の時の 5 段階に分類した．図 8 はユーザ A，図 9 はユーザ B のものを指す．x 軸は，5 分間での歩数量の度数分布を表している．x 軸が右に行くほど，よりオフライン活動を行った時間帯であることを表している．左側の y 軸は各時間帯のオンライン活動の量 (文章量) を示す．右側の y 軸は，各歩数の頻

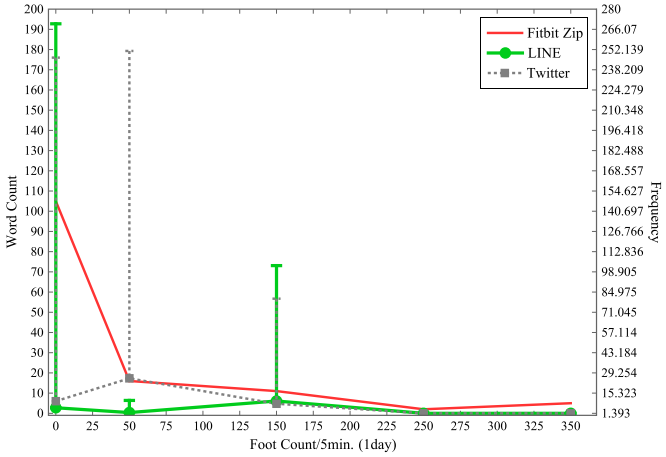

図 9　5 分間におけるオフライン活動量毎の度数分布とオンラインログの頻出頻度 (被験者 B)

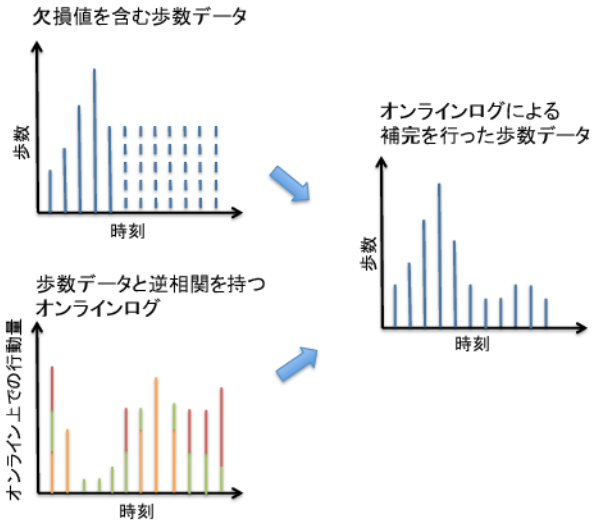

図 10　提案システムを用いた欠損値補完のイメージ図

度を示している．プロット点は，各オンライン活動による文字数の平均値を示しており，バーはその最大値と最小値を示している．殆どのプロット点が 0 に近い値にあるのは，LINE や Twitter の投稿におけるメッセージ数は基本的にとても少量であり，実験を通してその傾向が確認できた．

図 8，図 9 より全体的に，オフライン活動がされていない時間帯は，オンライン活動では長い文章の送信が行われるようになり，尚かつ平均値は他の時間帯と同様に低いことから，短い文章の送信は非常に多数行われていることが伺える．よってこの部分より，オフライン活動が控えめである時間帯では，オンライン活動は活発であるといえる．逆に，オフライン活動が多い時間帯では，オンライン活動が控えめになり，送信する文章の最大文字数が減少していることから，やはり移動中や活動中はオンライン活動を行うことが難しく，文章量といった活動の内容に制限がかかるのだと思われる．しかしながら，歩数が 200 以上である時間帯の活動は非常に少ないことから，オンラインログとオフラインログの相関関係を実証するためには，特にオフライン活動が活発な時間帯において，より多くのデータ数を確保する必要があると考えられる．

4. まとめ

本稿ではオンライン上で蓄積されるライフログと実世界での活動として蓄積されるオフラインログの抽出を行えるシステムを提案した．また，実際にシステムを構築し，得られたオンラインログが，オフラインログと相関関係を持っているかついて考察し，欠損値の推定，補完に有用であるかについて述べた．

実際に被験者から得られたオンラインログ，オフラインログを様々な視点から比較することで，両者には逆相関関係が見られた．今後の展望として，より多くのユーザから広い期間でデータを取得し，長い区間におけるデータの比較を行い，より統計的な視点からオンラインログ，オフラインログを観察し，それらの持つ特徴や相関関係の存在をより確固たるものとして証明したい．

これらのオンラインログや他のオンライン活動によって得られたデータが十分なものである場合，それらを欠損値補完へ応用することで，よりその日のコンテキスト情報に沿った行動推定が可能になると考えられる．例えば，Gmail 等のメールサービスや，Evernote[*10] 等による写真や文献の保存は，オフィス等の場では多く利用され，逆に動いている間の利用は少ないと考えられる．Facebook や Twitter 等における呟き機能の利用は，今現在の状況を報告する際に用いられるので，会社に留まらずあらゆる場所で行われる可能性があるが，文章を書く機能ということから，移動中よりもスポットに到着した時やスケジュールを終えた後に用いられるケースが多いと考えられる．Foursquare を代表するチェックインに関わるオンラインログは，外出中に，特に遠い距離の移動が行われた後に多く出現すると考えられる．図 10 は，提案システムを用いた欠損値補完アプローチのイメージ図である．

こうした各オンラインログが蓄積される裏側のユーザの行動には，それぞれ特徴があると考えられ，その時ユーザが何をしているのかというコンテキスト情報の推測に用いることができるのではと考える．

また，システム利用ユーザにもよるが，1 日の中でも比較的アクション量の多い LINE や Twitter における短いメッセージの送信は，他のオンラインログと比べて広い時間帯でデータを取得することができることが分かり，メールやメッセージ，ツイートのやり取りといったオンラインログは，オフラインログの推定にあたって非常に扱いやすいデータソースであると考えられる．今後は多くのユーザの間でもメール，メッセージの送信を扱うオンライン活動が，広い時間帯で取得可能なものであるかについても実証し，欠損

[*10] www.evernote.com

したオンラインログの推定に対してより有効なデータソースであるかについての考察を深めたい.

こうしたオンラインログ, オフラインログと実生活の関係を紐解くことの延長として, 健康支援に向けたサービスの範囲だけでなく, ユーザの健康状態, ストレス等の精神状態の推定や, それらを考慮した情報推薦, コンテキスト情報の把握といったユビキタスシステムの基盤にも用いることが可能であると考えられる. レシピ推薦等において, スケジュールからその日のコンテキスト情報を読み取る方法が考えられているが [11] , スケジュールの記入という能動的なアクションをユーザの負荷や, またスケジュールに書けない細かいオフライン活動を考慮できない点を考慮すれば, 暗黙的に取得できるオンラインログを用いた方法は有効であると考えられる.

参考文献

[1] T. Hashiguchi, H. Takeuchi, and A. Uemura. Highly advanced healthcare support services for the 21st century. *Hitachi Review*, Vol. 50, No. 1, pp. 2–7, 2001.

[2] 野口健一郎, 大谷真. Osi の実現とその課題. 情報処理, Vol. 31, No. 9, pp. 1235–1244, 1990.

[3] H. Takeuchi, et al. Automated healthcare data mining based on a personal dynamic healthcare system. Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE, 2006.

[4] N. Kawaguchi, N. Ogawa, Y. Iwasaki, K. Kaji, T. Terada, K. Murao, S. Inoue, Y. Kawahara, Y. Sumi, N.Nishio. HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings. ACM, Proceedings of the 2nd Augmented Human International Conference, pp. 27, 2011.

[5] ART. Donders, GJMG. van der Heijden, T. Stijnen, KGM. Moons. Review: a gentle introduction to imputation of missing values. Journal of clinical epidemiology. vol. 59, No. 10, pp. 1087–1091, 2006.

[6] P. Royston. Multiple imputation of missing values. Stata Journal, Vol. 4, pp. 227–241, 2004.

[7] J. Marchini, B. Howie, S. Myers, G. McVean, P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature genetics, Vol. 39, No. 7, pp. 906–913, 2007.

[8] Y. Nakano, H. Kawakami H. Tarumi. ライフログを共有する Life Networking Service. Interaction, 2012.

[9] http://www.softbank.jp/mobile/service/softbankhealthcare/

[10] 土方嘉徳. 情報推薦・情報フィルタリングのためのユーザプロファイリング技術. 人工知能学会論文誌, Vol. 19, No. 3, pp. 365–372, 2004.

[11] 三野陽子, 小林一郎. ユーザのスケジュールを考慮したダイエットのためのレシピ推薦. DEIM Forum, 2009.