情報処理学会論文誌　コンシューマ・デバイス & システム　Vol.10　No.1　15–25 (Feb. 2020)

コンシューマ・システム論文

自治体を軸とした業界横断型データ流通プラットフォームの開発

美原 義行[1,a)]　高山 雄策[2]　大部 隆二[3]　岩見 哲夫[4]　神山 肇[1]　松本 茂[4]　戸室 真敬[4]

受付日 2019年7月8日, 採録日 2019年12月2日

概要：本研究では，地域の課題やニーズをデータから把握することを目指し，地域経済圏のデータを業界を横断して集約して分析するデータ流通プラットフォームの開発を行った．我々は，札幌市をフィールドとして，観光分野での訪日外国人のさらなる集客に向けて，外国人観光に関わる市内約 30 社のデータを集約し分析を行った．扱うデータとしては，訪日外国人の購買データと宿泊データ，周遊施設における訪日外国人の入場者数データに加え，自治体におけるオープンデータと携帯電話基地局データ，その地域で発信された SNS データも集約した．本プラットフォームでは，同一業界内の全体傾向と差分が大きい点を自社の特異点として課題検出するだけでなく，特異点と相関の高いデータを業界を横断して検出することで，地域内で相互に連携した課題解決策の設計を支援することを目指す．本機能の実現に向け，購買データに関する特異点を自動で抽出する機能と，業界を横断したデータから特異点と相関の高いデータを抽出する機能，購買データの特異点抽出時に分析の横串となる商品分類を自動で付与する機能，それらデータが各施設から提供される際に個人を特定できない形に処理する統計加工機能の開発を行った．実際のデータ提供企業に特異点を提供し，気付いていなかった特異点も本機能により簡易に発見できることを確認した．

キーワード：データ流通，データ共有，特異点抽出，相関分析，第三者提供，統計加工

Development of a Cross-industry Data Aggregation Platform Centered on Municipalities

Yoshiyuki Mihara[1,a)]　Yusaku Takayama[2]　Ryuji Obu[3]　Tetsuo Iwami[4]　Hajime Kamiyama[1]
Shigeru Matsumoto[4]　Masahiro Tomuro[4]

Received: July 8, 2019, Accepted: December 2, 2019

Abstract: In this research, we aimed to extract regional problems and requirements from data generated in the city. We developed a data distribution platform that collects and analyzes regional data across industries. In cooperation with Sapporo City, we aimed to increase the number of foreigners visiting for sightseeing. We collected and analyzed data of about 30 companies in the city involved in foreign tourism. We collected foreign sales data, data on accommodation, number of visitors data in tourist facilities, open data in Sapporo City, GPS data using a mobile phone and SNS data transmitted in the area. In this platform, a point with a large difference between the index value of the company and the index value of the whole city is detected as a singular point. Furthermore, we aim to support the design of collaborative solutions within the region by detecting data highly correlated with singular points. To realize this function, we developed a function to automatically extract singular points of sales data, a function to extract data highly correlated with singular points, a function to automatically assign product classifications that are used in the analysis, and a function to process data statistically. We provided singular points to the companies. We confirmed that the singular points, which they did not notice, can be found easily by these functions.

Keywords: data distribution, data sharing, singularity extraction, correlation analysis, third party provided, statistical data processing

© 2020 Information Processing Society of Japan

1. はじめに

サービスを設計する際に，ターゲットとなるユーザの課題やニーズを把握することが一般的であるが，課題とニーズには地域性がある．たとえば，雪が多い地域における課題と雪が降らない地域では課題が異なる．ほかにも，有名な観光施設がある地域とない地域でも課題は異なる．我々は，地域のデータを集約して分析することで，その地域の課題やニーズを把握することを目指している．

地域経済圏の企業や団体が協力して，各組織のデータが集約され分析することで，その地域全体の傾向を把握できるようになり，マクロ的視点の分析が可能になることが期待できる．データを提供した各企業・団体においても，自組織のデータと地域全体の傾向とを比較することで，自組織の弱み・強みなどの特徴を把握できるようになり，ミクロ的視点の分析も可能となることが期待できる．自社データと，集約した様々なデータのなかから相関のあるデータを見つけることができれば，地域内での相互に連携した課題解決策を導くことも期待できる．

地域経済圏のデータを集約する先として，多くの企業のデータを集められるよう，協力を得やすい組織である必要がある．我々は，地域の中心である自治体が地域の企業の協力を得やすいという仮説を立て，自治体に地域経済圏のデータを集約するモデルを検討した．

我々は札幌市と連携（2015 年 9 月包括連携協定を締結 [1], [2]）し，札幌経済圏のデータを札幌市の設備に収集し分析する取り組みを行っている．取り組みの足がかりとして，急激に伸びている訪日外国人観光客（以下，インバウンド）の維持だけでなく，さらなるインバウンドの集客を目指し，観光分野での業界横断型のデータ収集・分析を行った．インバウンド観光客に携わる組織として，百貨店やドラッグストア，コンビニなどの商業施設，シティホテルやビジネスホテルなどの宿泊施設，公園や博物館などの周遊施設からデータを収集した．商業施設からはインバウンドの購買データとなる免税処理時のデータを収集する．宿泊施設からは宿泊データを収集し，周遊施設からは

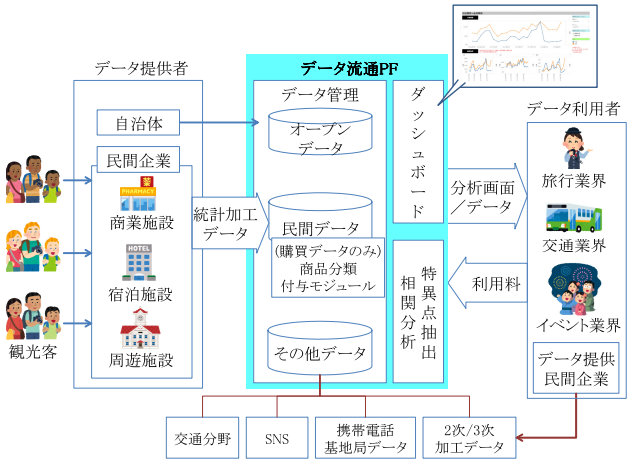


図 1　各施設のデータを集約して分析するデータ流通モデル

Fig. 1　A data aggregation platform that analyzes data for data providers.

入場者数データを収集する．これら施設からのデータに加え，地域に関する情報として自治体におけるオープンデータ[*1]や携帯電話基地局データ，その地域で外国語で発信された SNS データも集約した．携帯電話基地局データとは，携帯電話のネットワークを活用して作成される人口情報で，1 km メッシュ単位での国籍別の滞在人口を把握できる．

各施設のデータを集約して分析するデータ流通モデルを図 1 に示す．各施設のデータを集約する設備を，データ流通プラットフォーム（以下，PF）と呼ぶこととする．本 PF に業界を横断した様々なデータが集約され，集約したデータを可視化するダッシュボードも提供する．本取り組みに対して，札幌市内のインバウンド観光に関わる企業約 30 社（商業施設：6 社 369 店舗，宿泊施設：11 社 14 施設，周遊施設：11 社 13 施設）に参画いただいた．データ提供者以外のデータ利用者にも PF を通じてデータが流通し，活用されることも期待される．たとえば，交通業界の企業がメッシュ単位での人口データや商業施設・宿泊施設・周遊施設のデータから，バスの運行ルートの見直しを検討することも考えられる．また，旅行代理店などが同様のデータから新たなツアープランを設計することも考えられる．

まずはデータ提供者における課題の抽出と課題の解決を目指す．実現には以下の 3 つの処理が必要となる．

(1) 第三者提供に向けたデータ加工
(2) 統一的な商品分類の自動付与
(3) 特異点の自動抽出と特異点と相関のあるデータ抽出

企業のデータを自治体含めて外部に提供するときは第三者提供 [3] となる．したがって，企業はデータ提供時に個人情報を含まない形に事前にデータ加工を行い，PF に集約する必要がある（上記 (1)）．PF へのデータ集約後に，自社と全体傾向を比較して分析するためには，“横串” となるような分析軸が必要である．購買データにおいて横串と

[1] NTT 新ビジネス推進室
Strategic Business Development Division, NTT, Chiyoda, Tokyo 100–8116, Japan
[2] NTT コムウェア営業企画部 2020・地方創生ビジネス推進室
2020 Business Strategy Office, Marketing and Business Strategy Department, NTT COMWARE CORPORATION, Minato, Tokyo 108–8019, Japan
[3] NTT 西日本ビジネスデザイン部
Business Design Headquarters, NTT WEST CORPORATION, Osaka 540–851, Japan
[4] NTT 東日本ビジネスイノベーション本部 BBX マーケティング部
BBX Marketing Department, Business Innovation Headquarters, NTT EAST CORPORATION, Shinjuku, Tokyo 163–8019, Japan
[a)] yoshiyuki.mihara.cr@hco.ntt.co.jp

[*1] https://data.pf-sapporo.jp

© 2020 Information Processing Society of Japan

なる共通項目として，国籍や商品の分類がある．国籍の表現は統一的であるが，商品の分類については，各社独自の体系で管理されている．したがって，商品の軸で分析するためには，各社のデータを統一的な分類に振り直す必要がある．この作業をすべて手動で実施した場合，大量のデータを処理する人的コストが非常に高くなってしまう．したがって，商品の分類を自動で付与する必要がある（上記 (2)）．最後に，データ提供者が自社売上と全体傾向の差分が大きい点などの課題を簡易に把握するためには，PF 側で特異点を抽出して伝える必要がある（上記 (3)）．さらに，データ提供者側で特異点として見つかった課題や強みに対して解決する施策や伸ばす施策を簡易に設計できる必要がある（上記 (3)）．

上記 (1) に対し，事前の個人の同意なくデータを第三者に提供するため，我々は統計加工を行った．統計加工を行うことで，事前の同意なく第三者へのデータ提供が可能となる [3]．統計加工とは，データ内の複数の項目に対してグループ化を行い，値の平均化や合計化を行い，レコードをまとめていく処理である．グループ化した後は，同じデータが合計で規定数以上ない場合に，そのデータを削除する k-匿名化を行った [4]．この処理において，規定数以上の件数がない場合は削除するため，データの欠損が大きく発生する．しかしながら，購買額における国籍ランキングと，最も購買意欲が高い中国人の購買商品分類のランキングにおいて，統計加工前後での差は大きく発生しなかった．国籍別や商品別の分析において，実態に即したデータを残せることを確認できた．

この統計加工をデータ流通 PF で処理することは，企業内の未加工のデータが第三者に提供されたことになる．本研究では，上記の加工を行うプログラムを，各企業個別に開発し，それぞれに配布した．このプログラムにより，各企業で簡易にデータを加工することができ，データ流通 PF に統計加工後のデータを格納することができた．

上記 (2) に対しては，各社の統計加工後の購買データ内の商品に関する情報を入力として，その商品が属する商品分類を付与することを目指した．商品分類として，小売における商品が約 2,500 にまとめられた JICFS 分類 [5], [6] を採用した．ナイーブベイズ方式 [7] を用いて，JICFS 分類が付与された学習用データから，計算に必要な確率を事前に算出しておく．商品に関するデータが入力された際，形態素解析を行って単語分割を行い，各文字列に対応する事後確率を計算し，確率が高い JICFS 分類を出力する．手動で JICFS 分類を付与する際の誤り率も検証したことで，手動での精度と同等で，一部を自動化して付与コストを削減する運用方法も設計できた．

上記 (3) に対して我々は，売上分析のノウハウを持った識者による分析ロジックを，フローチャートに落とし込み，特異点抽出機能を開発した．この機能による各社の特異点抽出結果を商業施設 4 社に対して提示したところ，それまで自社内のデータ分析だけでは把握していなかった特異点が含まれていることを確認できた．本機能提供後，特異点をきっかけとして，各社において PF のダッシュボードの閲覧回数も増え，データを参照する足がかりとすることもできた．さらに，特異点を解決する施策の設計支援として，特異点と相関のあるデータを業界を横断して抽出することを行った．具体的には，経済活動に関連があると考えられる，月単位の合計降雪量データ，平均気温データ，イベント総件数データを市のオープンデータから抽出し，かつ，札幌市内で SNS に投稿された月単位の総件数，携帯基地局データから抽出したメッシュ単位での総人口数と，この特異点の相関を求めた．我々が抽出した，特異点と相関のあるデータを元にしたプロモーション活動は，2019 年 4 月時点で各社ともに実施されていないが，全商業施設から利用について前向きな意見をもらっている．施策設計に向けた足がかりとすることができた．

地域経済圏でデータを共有し，その地域全体の傾向を把握できただけでなく，その全体傾向と自社の傾向を比較できるようになったことで，自社の強みや弱みを把握できるようになった．地域によって様々な傾向が異なる．今回実証した札幌市は冬に向けて観光客が多くなるが，沖縄地方ではその傾向はあてはまらない．地域経済圏でデータを共有することで，正確に自社の弱み・強みを把握することが可能となる．業界を横断したデータから特異点と相関のあるデータが抽出されることで，地域内での企業間の相互の送客などの連携施策が生まれ，地域で一丸となってインバウンド集客を推進することが可能となる．

本稿の構成は以下のとおりである．2 章で関連する先行研究について述べ，3.1 節で自治体を軸としたデータ流通 PF に必要な機能をリストアップする．次に，3.2 節で要件を実現するためのデータ流通 PF のシステム設計を行う．4 章で評価について述べ，最後に 5 章で将来課題について述べる．

2. 関連研究

単体のデータだけでは実現できなかったことを，データを共有し合うことで実現する試みがなされてきた．文献 [8], [9] などは車載センサデバイスのデータを共有し合うことで，自車周辺の情報だけでなく，広範囲で周辺の混雑状況を把握できるようになる．文献 [10], [11] などは災害時に個人の移動実績データなどを共有し合うことで，避難場所までの移動可能経路を把握できるようになる．本研究は，上記のようなデバイス間，個人間のデータ共有ではなく，企業間でデータを共有することを特徴とする．

自治体を中心に企業が連携し，観光を盛り上げる試みとして，奈良県での観光アプリの取り組みがある [12]．この取り組みは，奈良県内約 550 店舗が参画し，それぞれが店

© 2020 Information Processing Society of Japan

舗情報を提示することで，コンテンツが充実した観光アプリが構築されている．我々の取り組みは，各企業における商業的な売上向上などを目指し，各企業の購買データや周遊施設の入場者数データを共有する．データを活用する側においても，業界を横断した様々なデータが集約されているため，ニーズや課題を多角的に抽出しやすくなることが期待される．

3. 自治体を軸としたデータ流通 PF の設計

3.1 設計における要件

企業や団体のデータを自治体に提供するときは，第三者提供となる．したがって，そのデータが氏名や住所，学歴，職歴などの個人情報を含むか否かが重要となる．個人情報を含む場合は，事前同意が必要となる．しかしながら，購買行為において，データの利用目的などを説明して同意をとるオペレーションは現実的ではない．個人情報を含まない形に変換し，かつ，元データへの復元が不可能な容易照合性をなくす形にすることで，個人の同意なく第三者提供が可能となる．商業施設と宿泊施設，周遊施設は，自社データを PF に格納する前にデータを加工する必要がある．これを要件 1 とする．

データ流通 PF へのデータ集約後は，商品分類ごとの傾向把握や，自社と全体傾向を比較するうえで，商品分類単位で比較を行うことが商業施設側の分析に必要となる．各社データにおいても商品の分類体系が含まれているが，各社独自の体系で管理されている．全社的な傾向を細分化して把握するためには，各社のデータを統一的な分類に振り直す必要がある．これを要件 2 とする．

次にデータ整形を行った後，各社にデータを可視化したグラフをデータ流通 PF で提供する（図 1 のダッシュボードモジュール）．可視化する画面としては，時系列ごとに全商業施設の売上データや，全宿泊施設の稼働率，全商業施設の入場者数データなどである．我々は，先行して数カ月分のデータに対して各施設のデータ整形までを行い，購買データに対して商品分類を手動で付与し，各施設のデータを可視化した画面を提供した（**図 2**）．

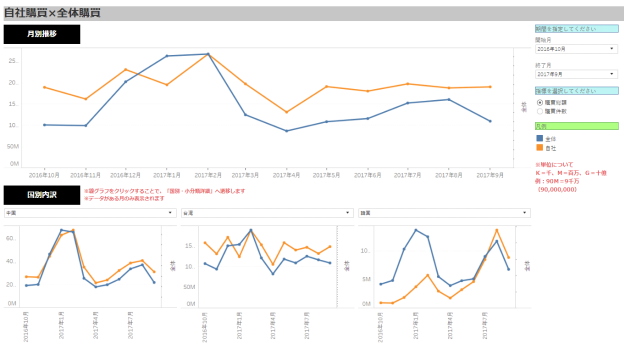


図 2　ダッシュボード例
Fig. 2　A dashboard example.

しかしながら，各企業の方々がデータから全体傾向を把握し，かつ，その全体傾向と自社データを比較して差分が大きいところを抽出して原因を深掘りすることができることは非常に稀であった．注目すべき点である特異点が自動的にレポートされれば，今まで気づくことができなかった特異点に簡易に気づくことが可能となる．したがって，特異点を自動的に抽出する必要がある．

特異点を抽出できた後に，その特異点を改善する，もしくは伸ばすためのプロモーション施策をデータ提供者側で設計しやすくすることが必要である．本取り組みでは，地域経済圏でデータを共有し合う．抽出した特異点と業界を横断して様々なデータの相関を求め，特に相関が強い説明変数を抽出することが可能となれば，地域経済圏内での企業間の連携も生まれ，相互に売り上げを伸ばすことも期待できる．たとえば，ある商業施設において特異点として抽出された国籍 A における商品分類 B の売上と，ある宿泊施設における国籍 A の宿泊者数データに相関が見られれば相互送客を行い，相互に売り上げを伸ばすことも期待できる．特異点抽出と相関分析を要件 3 とする．

3.2 要件実現に向けた設計

3.2.1 要件 1：第三者提供に向けたデータ加工

事前の個人の同意なくデータを提供するためには，統計加工する必要がある [3]．この処理により，元データへの復元が不可能となる．さらに，k-匿名化 [4] を行うことで，データの容易照合性を低減することが可能となる．

図 3 に購買データを用いた統計加工の例を記載する．日時を年月に汎化させたうえで，点線の領域でグループ化する．そして，同一レコードをまとめ，数量と全額を合計化し，グループ化した際のパスポート番号のユニーク数を件数として付与する．さらに個人を特定しにくくするため，特定確率を $1/k$ 以下にするよう，グループ化後（図 3）に k 人未満のデータを削除する，k-匿名化の処理を行う．

本取り組みにおいては，件数が 5 以上のデータのみを採

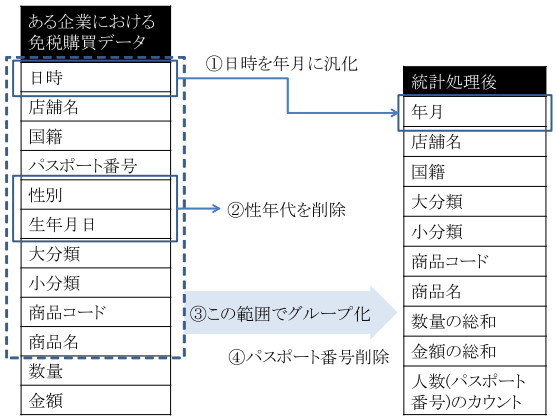


図 3　統計加工の例
Fig. 3　A flow of statistical processing.

© 2020 Information Processing Society of Japan

表 1　各データにおけるヘッダ
Table 1　Headers for each data.

購買データヘッダ
企業名
年月
店舗名
国籍
自社大分類
自社小分類
商品コード
商品名
共通大分類
共通小分類
売上数量
売上金額
購入人数

宿泊データヘッダ
企業名
年月
曜日
施設名
国籍
チェックイン数
宿泊人数
平均宿泊数
保有室数
売上金額

周遊データヘッダ
企業名
年月
施設名
国籍
入場者数

用（$k = 5$）することとしている．しかしながら，商品情報の粒度も細かく年代の粒度も細かいため，本処理を行うと件数が規定数以上になるレコードが少なくなり，データの欠損が多くなってしまう．

データの欠損を防ぐため，最低限必要なデータについてデータ提供企業にヒアリングした．国籍別のデータを収集できていない周遊施設以外の，商業施設と宿泊施設にヒアリングした．ヒアリングの結果から，商業施設・宿泊施設ともに国籍を軸にした売上分析が求められていたため，国籍を残すこととした．そして，時間軸の粒度としては年月で許容できることを把握できたため，年月単位とした．それ以外の性別・年齢はグループ化前に削除することとした．最終的な購買データ，宿泊データ，周遊施設の入場者データ，それぞれのヘッダ項目は**表 1** のとおりである．購買データは商品単位となっており，自社の分類と商品名が商品に関する情報となっている．宿泊データにおいては稼働率などの宿泊業界での様々な指標に対応できるよう，国籍ごとの宿泊人数，保有室数，金額が含まれている．

各社からの統計加工データは，システム構成を簡易にするため，CSV 形式のファイルとした．各施設は統計加工データである CSV ファイルをシステムへアップロードすることで，データを格納する．システムに CSV 形式でデータが投入されるため，簡易に既存データにマージできる．CSV ファイルとして，簡易に他企業のデータにマージできるよう，表 1 のとおり，静的なデータと動的なデータ含めてすべての情報を持つ設計とした．

グループ化する際の項目を粗くすることで，欠損を防ぐことはできるが，訪日人数が少ない国籍のデータが削除される可能性が高くなる．売上向上に向けて，新しい言語に対応してプロモーション活動を行うことや，建物内の案内板やパンフレットなどを用意することは非常にコストがかかる．したがって，現状は効果が高い国籍のみを把握できれば十分である．データ提供企業へのヒアリングにおいても，訪日人数が少ない国籍を把握する必要性が薄いことと，訪日人数が多く，プロモーション活動上重要な国籍に関する情報のみを把握できればよいことを確認した．

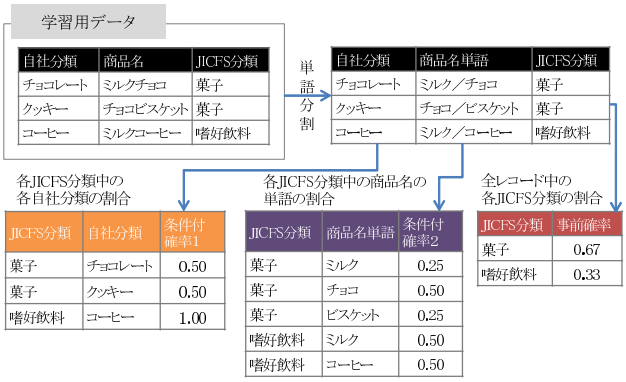


図 4　事前確率の付与例
Fig. 4　An example of calculating prior probabilities.

各社の生データのスキーマはまったく統一化されていない．そのため，まずはグループ化前の形式（図 3 左）として，PF に格納するスキーマに整形する．その後，グループ化し集計を行う統計化を行う設計とした．本処理をプログラム化して各商業施設と宿泊施設へ配布し，PF へデータを格納する前に実施していただく運用とした．

3.2.2　要件 2：統一的な商品分類の自動付与

各社の統計加工後の購買データ内の商品に関する情報を入力として，その商品が属する JICFS 分類を付与する．商品分類として，小売における商品が約 2,500 にまとめられた JICFS 分類 [5], [6] を採用した．JICFS 分類は，4 段階のレイヤで定義されており，基準書 [6] には各分類の定義が記載されている．たとえば，“インスタントコーヒー” という分類は，1 段階目が食品，2 段階目が飲料・種類，3 段階目が嗜好飲料，4 段階目がインスタントコーヒーとなっており，段階が深くなるにつれ粒度が細かくなっている．

本研究では，商品に関するデータに対して単語分割を行い，ナイーブベイズ方式 [7] を用いて，JICFS 分類が付与された学習用データから計算に必要な各確率を算出しておく．確率を計算する項目として，学習用データにおける各 JICFS 分類の出現頻度（事前確率）と，ある JICFS 分類における自社分類の文字列の発生確率（条件付確率 1），商品名の文字列の発生確率（条件付確率 2）がある．**図 4** の例を用いて説明する．まず学習データにおける自社分類，商品名について単語分割を行う．商品名に対しての条件付確率 2 は，分類 “嗜好飲料” は “ミルク” と “コーヒー” の 2 つの商品名単語と結び付いているので，“ミルク” と “嗜好飲料”，“コーヒー” と “嗜好飲料” が結び付く事前確率をそれぞれ 50%とする．分類と結び付く商品名の文字列の個数を分母として，分子を各文字列の出現回数とする（図 4）．“菓子” は “ミルク”，“チョコ”，“チョコ”，“ビスケット” の 4 つの商品名単語と結び付いている．“ミルク” と “菓

© 2020 Information Processing Society of Japan

購買データ

自社分類	商品名
クッキー	ミルクビスケット

自社分類	JICFS分類	条件付き確率1
クッキー	菓子	0.50
クッキー	嗜好飲料	0.01(該当なし)

商品名単語	JICFS分類	条件付き確率2
ミルク, ビスケット	菓子	0.25×0.25≒0.06
ミルク, ビスケット	嗜好飲料	0.50×0.01（該当なし）≒0.01

JICFS分類基準書を用いて事後確率を計算

自社分類	商品名	JICFS分類	スコア
クッキー	ミルクビスケット	菓子	0.67×0.50×0.06×1≒0.02
クッキー	ミルクビスケット	嗜好飲料	0.33×0.01×0.01×0.01(該当なし)≒0%

図 5　事後確率の付与例

Fig. 5　An example of calculating posterior probabilities.

子” が結び付く事前確率は 25%となるが，“チョコ” においては 2 回発生しているので 25% × 2 = 50%となる．自社分類における条件付確率 1 も同様の計算方法となる．事前確率である JICFS 分類の割合は，全レコードを分母として，各 JICFS 分類の発生数を分子とした割合である．図 4 では分類 “菓子” が 2 レコードあり，分類 “嗜好飲料” が 1 レコードなので，この割合は “菓子” が 67%，“嗜好飲料” が 33%となる．

これらの事前確率と条件付確率 1，2 をもとに，入力された自社分類と商品名の文字列に対して，各 JICFS 分類が結び付く確率（以下，事後確率）を計算する．**図 5** のように，自社分類が “クッキー” で，商品名が “ミルククッキー” が入力されたとする．まず自社分類の “クッキー” に対して，単語分割を行う．単語分割の結果，変わらず “クッキー” であるとして，“クッキー” と各 JICFS 分類の結び付く事後確率を求める．事後確率は，自社分類での条件付確率 1 と，商品名での条件付確率 2 それぞれに対して，JICFS 分類の割合（事前確率）とかけ合わせる．最終的にそれらをかけ合わせ，各 JICFS における事後確率を求める．図 5 の場合，自社分類 “クッキー” が JICFS 分類の “菓子” と結び付く条件付き確率 1 は 67% × 50% ≒ 33%となる．同様に，商品名の文字列に対して各 JICFS 分類との結び付く確率を求める．商品名の文字列を単語分割する．図 5 では “ミルククッキー” を単語分割すると，“ミルク” と “クッキー” に分割できたとする．この文字列が分類 “菓子” と結び付く事後確率を求めるため，“ミルク” と “クッキー” が “菓子” と結び付く，それぞれの条件付確率 2 とをかけ合わせる．ある文字列と JICFS 分類が結び付く事前確率が計算されていなかった場合は，非常に小さい数（図 5 の 0.01 のかけ合わせ）をかけ合わせることとした．最終的に，自社分類の条件付確率 1 と商品名の条件付確率 2，そして JICFS 分類の割合（事前確率）をかけ合わせて，JICFS 分類と紐づく事後確率を算出する．この際，最終的なかけ合わせの際に JICFS 分類基準書の定義内に一致する文字列が存在すれば，高い値の割合をかけ合わせる（図 5 の 1 のかけ合わせ）．

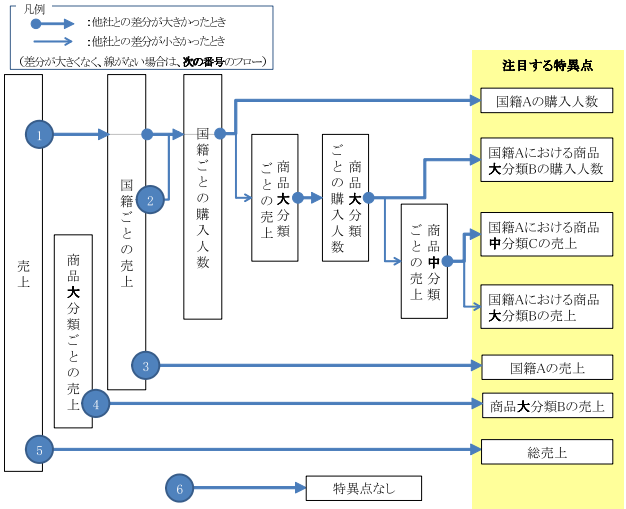


図 6　特異点抽出のフローチャート

Fig. 6　A flow for extracting points of interest.

本機能は，すでに割り当てられた組合せを優先的に検索する．入力となる商品情報の検索を行い，過去に JICFS 分類を付与済みの商品情報としてヒットした場合は，過去に付与した JICFS 分類をその商品情報に付与する．過去に JICFS 分類が付与していない商品情報が入力された場合に，各 JICFS 分類の事後確率を求め，適合すると考えられる JICFS 分類をスコア付きで出力する．このスコアは，全 JICFS 分類に対して事後確率を求めた後に合計が 1 となるように正規化した値である．出力としてスコア上位 3 位までの JICFS 分類を出力することとした．

3.2.3　要件 3：特異点の自動抽出と特異点と相関のあるデータ抽出

全体傾向と自社との差分が大きかった際に特異点として抽出する．特定の個社どうしで比較することは地域経済圏内での競争を生むことになり，地域内での相互連携を醸成する目的と異なる．比較としては，自社含めて，その地域内でデータを提供してくれた企業全体の平均と比較するように設計した．

本研究では，まずは購買データに対してのみ特異点抽出を行った．抽出する項目としては，売上金額と購入人数である．また，切り口として国籍，商品分類を用いた．その結果，自社が地域全体と比較して特定の国籍の特定の商品分類の売上が伸びている，もしくは落ち込んでいる，という傾向を把握可能とする．我々は，売上分析のノウハウを持った識者による分析ロジックを，フローチャートに落とし込んだ．具体的なフローチャートを**図 6** に示す．

基本的な考えとしては，まずは国籍単位での全体平均と比較する．そして，国籍単位での売上の差がある閾値以上であるとき，購入人数の変化による結果か，もしくは購入人数は変わらないが購買する商品傾向が変わり単価が変わった結果かを確認する．自社における購入人数の変化が

© 2020 Information Processing Society of Japan

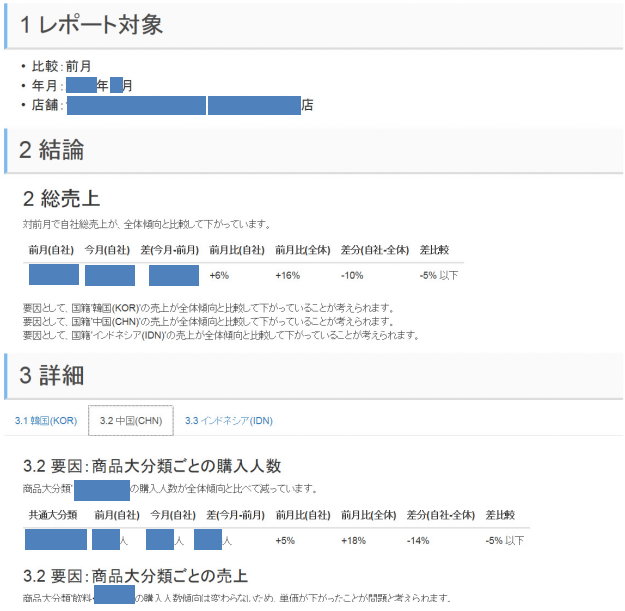


図 7　特異点抽出の画面例

Fig. 7　An example of a page displaying points of interest.

全社における変化と差が大きかった場合は，札幌全体としてインバウンドの人数が増えたのか，その店舗周辺のみで増えたのかを確認する．一方，購入人数において他社と差が小さい，つまり購買傾向が変わっていた場合は，売上が変わった商品分類を確認していく．上記のような項目に対して大きな差分が見えなかった場合は，条件を緩くしていくことで特異点を抽出する．

出力画面を**図 7** に示す．上部に店舗名と対象年月を表示する．メニューとしては 2 つあり，対前月における全体傾向との差分，対前年同月における全体傾向との差分を別ページで作成する．表示内容は概要と詳細の 2 つある．概要として売上全体傾向と国籍別の差分を表示する．詳細情報として，各国籍の売上が増減した要因を表示する．

特異点抽出後の相関分析においては，オープンデータや購買データ，宿泊データ，周遊データ，携帯基地局データ，SNS データを説明変数として設定できる．毎月発生する全商業施設の全特異点に対して，相関が高い最適な説明変数の組合せを求めることは，サーバリソース上不可能であった．本研究では，あらかじめ説明変数を定義しておき，抽出した特異点に対してその説明変数と重回帰分析を行い，相関を求めることとした．

今回事前に集約できた購買データは，2016 年 10 月から 2018 年 9 月までの 2 年分である．購買データの時間軸の粒度としては月単位であるため，特異点となる目的変数として 24 個のデータしか存在しなかった．説明変数の確定に向けて，モデルの検証を行う必要があったため，学習用の目的変数として 1 年分の 12 個用意し，説明変数としてこの数より少ない 10 個を選定した．特異点として見つかった課題を解決する施策を設計しやすいよう，時間に関する変数と，場所に関する変数，それ以外の変数に分けて説明変数の設計を行った．

表 2　特異点と相関分析を行う変数

Table 2　Variables for correlation analysis with points of interest.

カテゴリ	変数
場所	当月の都心エリアの周遊 2 施設の入場者数
	当月の郊外エリアの周遊 3 施設の入場者数
	当月の東部エリアの周遊 2 施設の入場者数
	当月の西部エリアの周遊 4 施設の入場者数
	当月の札幌駅周辺の滞在人口合計
時間	当月の札幌市の降雪量合計
	当月の札幌市の平均気温
	当月の札幌市内イベントの総件数
	前月の SNS の投稿数
その他	当月の札幌市内の宿泊施設総売上

表 2 の組合せは，検証用データを予測する検証を行い，平均的に精度が高かった説明変数群である．連携する相手，施策を行う場所が検討できるよう，周遊施設の場所を変数として設定した．施策を行うタイミングが検討できるよう，降雪量，平均気温，イベントの総件数，SNS の投稿数を設定した．天気との相関が見つかれば，天気が高い/低い，降雪量が多い/少ないときに施策を実施することで効果が上がることが期待される．イベントの総件数との相関が見つかればイベントが多いタイミングで施策を実施することで効果が期待される．SNS の投稿数との相関が見つかれば，投稿数を増やす施策や，投稿数が伸びているタイミングで施策を実施することで効果が上がることが期待される．宿泊総売上は，本来企業個別に相関を見い出すべきであるが，変数の個数の制限により企業個別のデータを説明変数とできなかった．これらの説明変数と特異点の相関分析を自動で実施する．

4. 評価

本システムの評価を要件とともに行っていく．

4.1 統計加工におけるデータ欠損の影響（要件 1）

要件 1 の実現においては，データの欠損が発生する．この欠損においてデータの信頼性が崩れることが懸念される．そこで，本機能の評価として，実際のデータとの乖離を検証する．本データにおける目的は，売上額などの絶対値よりも，ランキングなどの相対的な傾向の把握である．事業者に把握したい傾向をヒアリングし，以下の検証項目を選定した．ある店舗において，2 カ月分の統計加工前データに対してランキングを作成してもらい，5 位までの順位が合致しているか否かを検証した．

- 購買額における国籍ランキング（1 位から 5 位）
- 中国人の購買商品分類ランキング（1 位から 5 位）

© 2020 Information Processing Society of Japan

表 3　統計加工処理の有無によるランキング結果の差異 1
Table 3　1st difference in ranking results depending on the presence or absence of statistical processing.

購買額における国籍ランキング	統計加工処理なし		統計加工処理あり	
	国籍	金額の割合 (%)	国籍	金額の割合 (%)
1	A	59.1	A	60.3
2	B	34.1	B	37.3
3	C	2.8	C	1.0
4	D	2.6	E	0.8
5	E	1.4	D	0.6

表 4　統計加工処理の有無によるランキング結果の差異 2
Table 4　2nd difference in ranking results depending on the presence or absence of statistical processing.

中国人の購買商品種別ランキング	統計加工処理なし		統計加工処理あり	
	商品種別	金額の割合 (%)	商品種別	金額の割合 (%)
1	a	41.3	a	40.3
2	b	17.7	b	17.9
3	c	15.4	c	15.4
4	d	13.3	e	13.4
5	e	12.3	d	13.0

結果を**表 3**，**表 4** に表示する．店舗のプライバシーのため国籍や商品分類はマスクしている（ただし購買金額 1 位の国籍は中国である）．どちらのランキングにおいても 4 位の項目と 5 位の項目が逆となった．ただし，購買額における国籍ランキングでは，統計加工処理のなしとありで，金額の割合ベースでは，4 位と 5 位で 2 pt 程度の差でしかない．同様に，中国人の購買商品種別ランキングにおいても，統計加工処理のなしとありで，金額の割合ベースでは，4 位と 5 位で 1 pt 程度の差でしかなかった．傾向を維持できていることを確認できた．我々はこの結果を商業施設にフィードバックしたところ，2 pt 程度の差は施策の設計上問題とならない精度であることも確認できた．

統計加工処理では，データの欠損が発生しているため，処理の前後で合計額が異なる．しかしながら，商業施設におけるインバウンドへのプロモーション活動は言語対応のコスト的障壁が高いため，多くの国（言語）に対応できていない．現状は，5 カ国程度の情報で余りある状況であり，札幌市の全体傾向把握として，上位 5 カ国の売上比率が重要となることも確認した．

4.2　統一的な商品自動分類付与の正解率（要件 2）

要件 2 の検証においては，JICFS 分類の付与精度を検証し，運用方法の設計につなげる．この分類の精度が低い場合は，実サービスとして利用することができない．精度検証として，5 社分データの商品に関する文字列 13,689 件（重複なし）に対してクロスバリデーションを用いて検証を行った．**表 5** に認識精度を示す．精度は 78.8%であった．JICFS 分類において，大分類は 31，中分類は 362，小分類は 2,489 ある．中分類レベルの精度であれば，89.1%となった．また，候補として 3 つまで JICFS 分類を提示するが，3 つの中に正解が入っている確率は 89.0%となった．

表 5　JICFS 分類の自動付与精度 (%)
Table 5　JICFS classification prediction accuracy (%).

	候補 1 一致	候補 1〜3 のいずれかが一致
大分類（JICFS 分類 2 段階目）	94.7	97.9
中分類（JICFS 分類 3 段階目）	89.1	95.0
小分類（JICFS 分類 4 段階目）	78.8	89.0

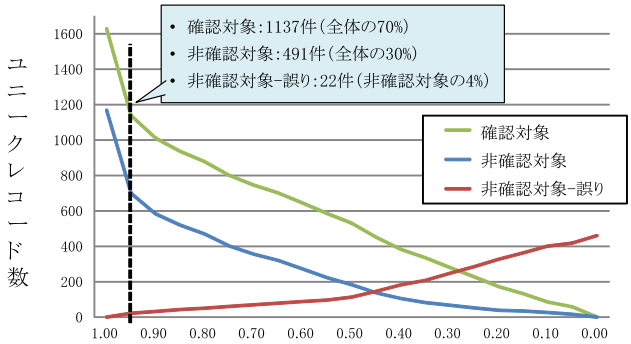


図 8　スコア閾値と確認対象件数の推移
Fig. 8　Changes in score threshold and number of items to be confirmed.

この機能を用いて手動での付与コストを削減するためには，手動での付与精度と同等の精度であることが求められる．我々は実際に手動で付与した際の精度を検証した．学習用データの JICFS 分類は手動で付与していたが，その学習用データをサンプリングして，正確な JICFS が割り当てられていたか否かを確認した．正解データを付与した人員と，サンプリングしたデータを再確認する人員は別であった．2,579 レコード（重複なし）のデータに対して再確認したところ，102 レコード（4.0%）に誤りがあることが判明した．誤りとしては，以下の種類があった．なお，本誤りを修正し学習した際の検証結果が表 5 である．

- 分類基準に詳細に内容が記載されていたが，その基準書の見落としにより類似の分類に割り当てられる誤り
- 同じ種類の商品（たとえば内容量が異なる程度の違い）を異なる分類に割り当てる誤り

正解データを付与した人員は複数人であったため，付与ポリシの統一が図れていなかったことが原因と考えられる．この結果から，手動と同じ精度で判定できている部分を自動化し，それ以外を手動で判定することで，手動での付与コストを削減することが可能となる．**図 8** は本機能が出力するスコアと精度の関係である．

“確認対象” の線は，横軸のスコアを閾値とした際に，その閾値以下となる件数を示した線である．閾値を高くすればする一方で，分類の付与結果を正しいと判断せずに手動での確認を行う件数（“確認対象”）は増加する．“非確認対象” の線はその逆で，横軸のスコアを閾値とした際に，閾

© 2020 Information Processing Society of Japan

表 6　1 カ月における 1 店舗あたりの特異点の平均数
Table 6　An average number of points of interest per store in a month.

企業 A	企業 B	企業 C	企業 D
4.0	9.0	3.3	13.8

表 7　把握できていなかったと回答を得た特異点の割合
Table 7　A percentage of points of interest that were answered that they could not be grasped.

企業 A	企業 B	企業 C	企業 D
42.9%	88.7%	9.1%	61.8%

値以上となる件数を示した線である．閾値を高くすればするほど，分類の付与結果が正しいと判断し，手動での確認を行わない件数（“非確認対象”）は減少する．“非確認対象–誤り” は，スコアが閾値以上となり，手動での確認が不要と判定した商品の情報に対して，付与した JICFS 分類が誤っていた件数を示した線である．スコアが高くなればなるほど正解率は上がるため，誤り率は減少する．

手動付与の誤り率が 4.0%であったため，グラフから自動化でも同じ精度を維持できるスコアは 0.95 であることが分かる．スコア 0.95 以上のレコードを自動登録させ，それ以下のスコアを手動での付与処理に回すこととした．本機能により，手動で付与する精度を維持しながら，約 30%（図 8）確認するレコードを減らすことが可能となった．

4.3　特異点の自動抽出の意外性の主観評価（要件 3）

意外性の主観評価を行い，把握できていた特徴か否かを商業施設 4 社からヒアリングした．ヒアリングは 2019 年 4 月に行い，2018 年 10 月から 2019 年 1 月の購買データに対してヒアリングを行った．

各社の 1 月の 1 店舗あたりの特異点の平均数は**表 6** のとおりである．図 7 における，全体の売上や国籍ごとの売上，その要因も 1 つとして数えた数字である．決して多すぎる数ではなく読みきれない量ではないと判断できる．特異点の出力は店舗単位であり，店舗数が非常に多い企業も存在する．各企業に対して 2 カ月分，1〜2 店舗分程度の特異点を抽出し，自社内データの分析などから把握していた内容か否かをヒアリングした（**表 7**）．全企業において少なからず 0%ではなく，把握できていなかった特異点が存在した．その要因としては，大きく以下の要因があった．

- 主要な売上を占める中国人のみの動向のみを把握し，それ以外の国籍の傾向を把握していない．
- インバウンドの売上において商品分類単位での集計をしていない．

中国人以外の特異点が出ていた企業や，商品分類まで深掘りした特異点が出ていた企業において，表 7 の値が高くなっている．データのみからは見つけることが困難であった特異点を自動で簡易に抽出でき，本機能が有効であることを確認できた．本画面を 2019 年 3 月末から提供したところ，2019 年 4 月までの 1 カ月で PF のダッシュボード画面への月間アクセス数が，倍以上に向上したことからも有効であることを確認できた．データ提供企業である一社が，特異点抽出において負の方向に他社との差分が大きかった国籍に対して，現地での日本観光博に出展し，その施設のプロモーション活動を行ったところ，売上が 3 倍以上になった，という結果も現れた．

特異点と相関があるデータを抽出する機能に対して，執筆時点の 2019 年 4 月時点で本機能を活用した施策はまだ実施されていない．本機能の有効性について各商業施設にヒアリングしたところ，全社から施策の設計に使っていきたい，という意見をいただいた．本機能から導き出される相互連携案に対しても，実行に向けて非常に前向きな姿勢であった．本機能が，施策設計の足がかりとなることを確認できた．説明変数の粒度が粗いため，個別の企業間の関連の把握まで至ることはできなかった．相関分析を様々なデータに対して行えるよう本取り組みを継続し，データ量を増加させていくことが重要となる．

5.　将来課題

1 つの都市で本 PF を実現したが，他都市へ展開することで分析の幅も広がると考えられる．参画都市全体で平均化して，その全体傾向と自都市の傾向を比較することで，地域固有の特異点も見えるようになることが期待できる．本研究で設計・開発したデータ流通 PF を今後，他都市へ展開していきたい．

本研究では特異点に対する施策設計の支援として，相関分析を行った．2 つの企業間のデータに相関が見つかった際は，その企業間での相互連携を行うことで，2 つの企業ともに売り上げが向上することも期待できる．一方で，個人の動線を追って企業間の関連を導き出す手法も存在する．今回対象としたデータは，インバウンド観光のデータであり，購買施設や宿泊施設の元データにはパスポート番号が含まれている．パスポート番号で各データを結合させることで，施設間の関連を把握することが可能になる．個人情報保護法においてパスポート番号は個人情報として扱われる [3]．したがって，パスポート番号を活用した分析を行うためには，法改正が必要となる．秘密計算 [13], [14] を活用できれば，各施設におけるインバウンド観光の各データをパスポート番号をキーに安全に結合していくことが可能となる．各施設間の結び付きが把握できるようになれば，さらなる地域経済圏内での企業間の相互連携が深まることが期待できる．今後，法改正含めて，秘密計算を活用した分析も視野に入れていく．

6.　おわり

本稿では，自治体を軸としたデータ流通 PF について述

© 2020 Information Processing Society of Japan

べた．地域経済圏のデータを自治体を軸に集約し，地域の全体傾向だけでなく，その全体傾向と各社のデータを比較することで企業ごとの特異点を把握できるようになった．特異点と相関の高いデータを業界を横断して検出することで，地域内で相互に連携した課題解決を簡易に設計できるようになった．本取り組みにおいて，各社のデータを第三者のデータ流通 PF に集約する際に，事前の個人同意が必要となくなるだけでなく，データの傾向も変わらない統計加工処理を行った．かつ，分析する軸となる商品分類を自動で付与できるようになった．今後は，観光だけでなく他の分野に関するデータを集約・分析し，地域内のニーズと課題の抽出に貢献していきたい．さらに，他の都市地域にも拡張し，地域間での比較を可能にすることによって，地域全体の課題をより浮き彫りにさせることも目指していく．

謝辞　本研究は，経済産業省「平成 29 年度補正予算 産業データ共有促進事業費補助金」を活用した「業界横断型の官民データ利活用プラットフォーム事業」（さっぽろ産業振興財団実施）の成果を含んでいる．

参考文献

[1] 札幌市：日本電信電話株式会社との協定（オンライン），入手先 ⟨http://www.city.sapporo.jp/shimin/support/partner_ntt/ntt_top.html⟩（参照 2015-09-16）．
[2] 日本電信電話株式会社：札幌市と日本電信電話株式会社との「さっぽろまちづくりパートナー協定」について（オンライン），入手先 ⟨http://www.ntt.co.jp/news2015/1509/150916a.html⟩（参照 2015-09-16）．
[3] 個人情報保護委員会：個人情報の保護に関する法律についてのガイドライン（通則編）（オンライン），入手先 ⟨https://www.ppc.go.jp/files/pdf/guidelines01.pdf⟩（参照 2016-11-30）．
[4] 長谷川聡，正木彰伍，岡田莉奈：大規模データを実用的な速度で処理可能な匿名化ライブラリの設計と実装評価，コンピュータセキュリティシンポジウム 2017 論文集，Vol.2017, No.2 (2017).
[5] 一般財団法人流通システム開発センター：JICFS 分類コード一覧（オンライン），入手先 ⟨http://www.dsri.jp/database_service/jicfsifdb/data/1101jicfs_bunrui-ichiran.pdf⟩（参照 2010-01-14）．
[6] 一般財団法人流通システム開発センター：JICFS 分類基準書（オンライン），入手先 ⟨http://www.dsri.jp/database_service/jicfsifdb/data/1312jicfs_bunrui-kijyunsho.pdf⟩（参照 2013-12）．
[7] Drew Conway, John Myles White.：入門 機械学習，オライリージャパン (2012).
[8] 藤田　敦，梅津高朗，山口弘純，東野輝夫，金田　茂，高井峰生：車車間通信を用いた車両間協調による周辺車両群の存在把握，情報処理学会論文誌，Vol.56, No.11, pp.2092–2105 (2015).
[9] 佐合弘行，篠原昌子，原　隆浩，西尾章治郎：車車間通信を用いた車両間協調による周辺車両群の存在把握，情報処理学会論文誌，No.78, pp.115–122 (2006).
[10] 黒島理礼，吉木大司，森　信彰，松本佳昭，亀川　誠，藤川昌浩，松野浩嗣：地域コミュニティ活動と連携した被災情報提供システムの開発，研究報告モバイルコンピューティングとユビキタス通信（MBL），No.7, pp.1–6 (2013).
[11] 服部聖彦，大和田泰伯，加川敏規：端末間のすれ違い通信を用いた避難経路情報の共有と群衆移動の大域的推定，研究報告モバイルコンピューティングとパーベイシブシステム（MBL），No.5, pp.1–4 (2015).
[12] 林田平馬，増山史倫：観光行動情報収集のための観光案内アプリの開発と実証～地域情報の持続的なデジタル化に向けて～，研究報告情報システムと社会環境（IS），No.8, pp.1–6 (2018).
[13] Kikuchi, R., Chida, K., Ikarashi, D., Ogata, W., Hamada, K. and Takahashi, K.: Secret sharing with share-conversion: Achieving small share-size and extendibility to multiparty computation, IEICE Trans., Vol.98, No.A(1), pp.213–222 (2015).
[14] 桐淵直人，五十嵐大，諸橋玄武，濱田浩気：属性情報と履歴情報の秘匿統合分析に向けた秘密計算による高速な等結合アルゴリズムとその実装，コンピュータセキュリティシンポジウム 2016 (2016).



美原 義行　（正会員）

2006 年日本電信電話株式会社（サービスエボリューション研究所）入社．以来，IoT ネットワーク管理サービスの研究開発，IoT ネットワーク管理プロトコルの標準化に従事．現在，日本電信電話株式会社サービスエボリューション研究所主任研究員．博士（情報学）．



高山 雄策

2009 年 NTT コムウェア株式会社入社．SFA/CRM システムの開発，およびデータ分析によるユーザ動態調査やマーケティング支援業務，新規ビジネスの立ち上げに従事．現在，NTT コムウェア株式会社テレコムビジネス事業本部 SO ソリューション部スペシャリスト．



大部 隆二

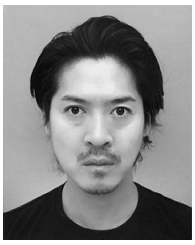
1999 年日本電信電話株式会社（同年 7 月に西日本電信電話株式会社設立に伴い移行）入社．以来，法人営業を経てサービス開発業務，人材育成業務等に従事．現在，西日本電信電話株式会社ビジネスデザイン部担当課長．

© 2020 Information Processing Society of Japan



岩見 哲夫

1999 年日本電信電話株式会社に入社．同年 7 月から東日本電信電話株式会社にて初中等，高等教育における教育 ICT イノベーションプロジェクトに従事．2012 年福島支店法人企画を経て，2015 年復興庁へ出向し震災復興業務に従事．現在，ビジネスイノベーション本部第三バリュークリエイト部第二バリュークリエイト（大学担当）担当部長．



神山 肇

2002 年株式会社 NTT データ入社．以来，環境エネルギー，ヘルスケア，リテール分野での新規事業創出に従事．現在，株式会社 NTT データ ITSP 事業本部 SDDX 事業部担当課長．



松本 茂

1998 年日本電信電話株式会社に入社．同年 7 月から主に金融セクターのネットワーク統合プロジェクト等に従事．2008 年東日本電信電話株式会社研究開発センタにて設備の維持管理計画策定業務に従事．2014 年より同社ビジネスイノベーション部にて法人営業に従事．2018 年より 1 年間さっぽろ産業振興財団特別研究員（委嘱）．現在，千葉事業部 BI 部テクニカルソリューショングループ担当部長．



戸室 真敬

2000 年東日本電信電話株式会社入社．東京支店にて民間系の法人ユーザに対する営業活動に従事．その後，NTT 出版株式会社，株式会社 NTT-ME の経営企画業務等に従事．現在，株式会社 NTT 東日本関信越栃木支店ビジネスイノベーション部にて栃木県内の民間系の法人ユーザに対する営業活動に従事（担当部長）．

© 2020 Information Processing Society of Japan