博士論文

ユーザを介した複数公共データの統合による知識抽出に関する研究

指導教員 坂田 一郎 教授

東京大学大学院 工学系研究科 技術経営戦略学専攻

大知 正直

2017年1月17日 提出

博士論文　2016 年度（平成 28 年度）

ユーザを介した複数公共データの統合による知識抽出に関する研究

要旨

公共データの公開及び活用は, 政府が 2016 年秋に設置した「未来投資会議」においても重要なテーマとして挙げられ，我が国の経済成長や社会課題の解決に貢献するものとして，重点的な推進が企図されている．公共データはこのように社会的にも大きな関心が寄せられている一方，現時点では限定的な利用にとどまっている．さまざまな公共機関，事業者が持つ公共データを，相互に提供しあうことによって, より大きな付加価値を引き出そうとしないのはなぜだろうか．本研究は，このような問いから始まっている．本論文では，公共データの活用が進まない現状に対する課題を，複数公共データの統合・活用に関する技術的な課題及びデータ保有者間の協力に関する障壁が存在していることにより，複数のデータを分析することで得られる知見の価値が，提供するリスクと比較し低くなっているものと捉え，技術的な課題の解決と障壁の低減により，リスクを上回る価値のある知見の抽出が可能となり，複数公共データの統合・活用が進むという仮説の検証を行っている．

そもそも複数のデータを持ち寄り，分析しようとする動機は何だろうか．それは各データを持ち寄り提供しようとする人々（公共機関，事業者）が分析をしようとする対象に何らかの共通項があると考えるからであろう．本研究では，分析対象となるデータ群に共通する何らかの対象を単にユーザと呼ぶ．そして，このユーザに関する情報がどのように保存されているかは，データの保有者によって内容，形式ともに多様である．中には実際は，全くユーザが利用していない無関係なデータもある可能性がある．そこで，本論文ではユーザを中心としてデータを横断して分析しようとするのではなく，データソースを中心に分析を行う異種データ間のユーザ移動ネットワークという考え方を提案する．そして，データソース間を移動するユーザを介した複数公共データの統合による知識抽出を試み，その実効性や有用性につき検証を行う．もちろんデータソース間の保存内容，形式に差異が大きい場合，通常，知識抽出は困難になるため，データの形式を想定することで，それぞれの場合における知識抽出法の有効性に関する検証を行った．具体的には，データソース間で

データを生成したユーザを一意に特定できる場合と，一意には特定できないが部分的に特定可能な場合である．ここで部分的に特定可能であるとは，データ間を移動するユーザ群の持つ共通の属性，行動等で集団として特定可能な場合を指す．そして，それぞれの場合において，有効な知見を獲得するための分析の枠組みを示した．複数公共データの利用又はデータ提携に関しては，技術的な課題以外に，データ保有者間の協力に関して，情報の守秘，便益の不均衡といった課題が存在する．そこで最後に，実務的な仮説の検証を行うために，前章までで開拓した技法群の考え方を参考とした上で,それぞれの事業者の感じるリスクを解決しデータ提供を促すための調停機能を実装した独自のシステムを開発し，実際に各事業者の分析担当者の利用に供した上で評価を受けた．これらの検証を通じて，データを組み合わせて分析することで得られる知見の価値がデータを他社に提供するリスクを上回れば，データによる事業者間提携が促進されることを明らかにした．

複数のデータソースから得られたデータを組み合わせて分析するための統一的な枠組みは存在せず，また，実社会においてもデータを組み合わせて分析を行う取り組みはあまりなされていない．一方で，複数公共データの統合による知識抽出は，事業者側の需要は大きいと考えられる．第1章では，多様なユーザが日々大量にデータを生み出しており，そのデータはプラットフォームを提供している各公的機関，事業者で分散されて保存されているが，この分散保存されているデータを集め，統合して分析するにはいくつかの課題があるために，有効に活用されていないことを説明する．そのような現状認識を踏まえ，本論文では，データ分析によって得られる知見の価値が低くなっていること，企業がデータを提供することにリスクを感じていることの２点が課題であることを示す．本論文では，これらの課題の解決方法として,「ユーザを介した複数公共データの統合による知識抽出」手法を提案する．そのためにまず，獲得する知見の価値を高めるための方法の一つとして，異種データ間のユーザ移動ネットワークの提案を第3章で行う．この際，各公的機関，事業者間の持つデータの性質によって２つの場合に分ける．具体的には，データソース間でユーザを一意に特定できる場合と部分的に特定可能な場合とで，具体的な知識抽出のための分析を行うためのそれぞれのユーザ移動ネットワークの枠組みについて議論する．

第4章では，データソース間でユーザを一意に特定できないが部分的に特定可能な場合における知識抽出手法の提案とその有効性の検証を行った．ここでは，データ間で共通するユーザの興味，行動という点でユーザ移動ネットワークを構成する．このユーザ移動ネットワークを用いて，ヒット商品におけるユーザの検索行動が複数のメディアでの露出にどのように影響されているか，という分析を行っている．ここでは，人々の関心というものはメディアでの露出量と大きな相関があり，メディアで露出すると一定の割合でユーザの興味が喚起され検索行動を行うという仮定を

おくことで，複数のデータソースをネットワーク化している．ソーシャルメディア上での関心の盛り上がりを「口コミ指数」として指標化することを提案し, ソーシャルメディア内における関心の盛り上がりが, 実際のヒット現象に与える影響を明らかにした．

次に第 5 章では，データソース間でユーザを一意に特定できる場合における分析手法の提案とその有効性の検証を行った．本章では，それぞれのユーザによるデータ間の移動の系列によって，ユーザ移動ネットワークを構成する．具体的には，公共交通の複数の運営会社から提供された乗降記録を用いて，ユーザの移動目的を推定する分析手法（「移動目的推定モデル」）を提案している．それにより特に，ユーザを特定できる移動ネットワークにおいては，文脈情報の利用を積極的にすることが移動目的の推定に重要な役割を果たしていることを示した．また, 人々は, 用事はなるべく近くで済まそうとするが，重要な目的がある場合には遠方へ移動する，等の知見を定量的な形で抽出することに成功した．

第 6 章では，事業者間のデータ提携における課題，利点を明らかにするため，前章までに示した分析手法の考え方を参考とし，独自の調停機能等とを装備したシステムを開発し，その有効性と有用性に関して検証を行っている．実際に，複数の公共交通の運営会社からデータの提供を受け，それらを統合し，各社局の担当者にとって興味を持つ分析を自由に行えるように設計を行った．その際に，それぞれの会社の持つ分析したい内容に対する希望と他社に開示したくない情報を把握した上で，便益と不利益の調整を試みている．実際の利用の後のアンケート調査を分析した結果，構築したシステムの有効性と，それぞれのデータ提供者が不利益にならないような設計 (調停者の存在) が重要であることがわかった．

本論文では, 複数公共データの統合・活用に関する技術的な課題及びデータ保有者間の協力に関する障壁を乗り越えるため，分析の枠組み，手法及び調停機能等を実装したシステムの提案とその有効性等に関する検証を行った．複数公共データの統合による知識抽出という複雑な課題に対し，一定の枠組みを示したことで，実際の事例において知識抽出の手法の開発，選定に注力することが可能になったと考えている．今後，本論文の提案手法を用いることで，データを複合的に組み合わせて分析することによって得られる知見の価値が，自社のデータを他社に提供するリスクを上回ることについて社会で認知されるようになれば，データの事業者間提携がこれまでよりも促進されることが期待できる．

東京大学大学院 工学系研究科 技術経営戦略学専攻

大知 正直

目 次

図 目 次

表 目 次

第1章　序論

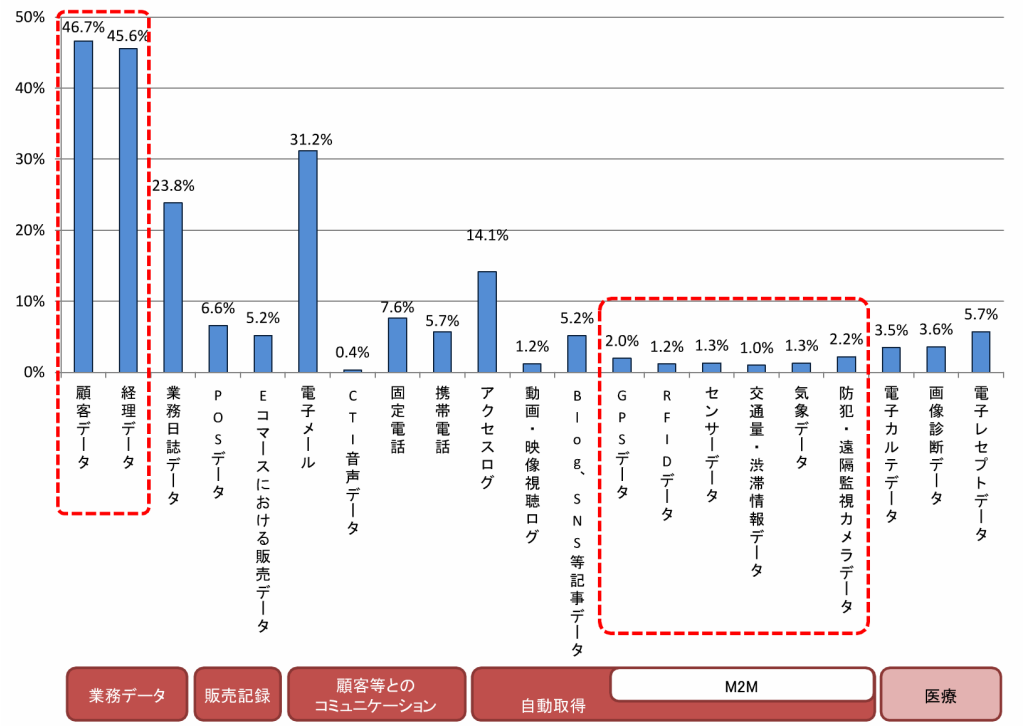
企業の分析担当者が，ユーザの行動をより詳細に理解するためには複数のデータを組み合わせて分析した方が良い，と考えるのは自然なことであろう．近年ではIoT化の流れにより多くの種類のデータが蓄積されるようになったが，各データ保有者の企業間連携における問題解決と，異なる種類のデータを結び付けて知識を抽出する手法の開発は進んでいない．特に政府は，公共データのオープン化とその活用を積極的に推進しており，複合的なデータ分析手法の確立の要請は大きい．そこで，本章では以上の背景について詳細に説明し，問題意識，目的を明らかにした上で，本研究で提案するユーザ移動ネットワークに着目して異種データを統合し，知識を抽出する手法について説明する．そして，提案手法の検証結果の概略をもとに本研究の貢献を説明する．

1.1　研究の背景

1.1.1　官民データ活用の推進とデータ流通の拡大

政府は官民データの活用を強く推進している．2014年にサイバーセキュリティ基本法の制定，2015年に個人情報保護法の改正を行い，個人データの保護と活用について，国及び地方公共団体の責務を明らかにした．そして，2016年に官民データ活用推進基本法を制定し，行政手続きの原則ＩＴ化を定め，公共データのオープン化を強く推進することで，公共データの生成，流通，共有，活用の飛躍的拡大を狙っている．特にこのような公共データの官民での活用は，2016年に政府によって設置された「未来投資会議」においても度々検討されている．特に公共データの徹底的な開放，活用は，大きな議題として位置づけられ，今後2020年を目処に集中的に推進していくとしている．

このような政策的な背景の元で，公共データと民間データを組み合わせ，どのように分析を行い，知識を抽出するかは大きな課題である．



※母集団は、何らかの分析をしていると回答した企業等（n=3, 357）

図 1.1: 各データを活用している企業等の割合 (「ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究 (2015)」から引用 [85]).

1.1.2　異種データの増加と官民のデータ活用

実社会にあるあらゆるデバイスがインターネットに接続されるモノのインターネット化 (Internet of Things: 以下，IoT[5]) という現象が進んでいる．この IoT 化が進行するにつれて，さまざまなデバイスから大量のデータが開発元サーバーへ送信され蓄積されるようになってきている．しかし蓄積されたデータが十分に分析され，意思決定に活用されているとは言えない．例として，国内のデータ活用状況を図 1.1 に示す．これによると，顧客データや経理データなどの業務上必要なデータは半数近い企業で分析に活用されているが，今後 IoT 化が進むに連れて，増加が想定される GPS データ，その他センサーデータ，交通量データ，気象データなど，自動的に蓄積されていくようなデータはほとんど活用されていないことがわかる．

一方で，この活用されていないデータは現状でどのように分析されているのであろうか．総務省発行の「ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究」によると，あまり活用されていなかった交通量データや気象データを分析する場合には，半数以上の企業で，単一のデータのみで分析するのではなく，複数のデータを組み合わせて分析していることを示している．これは交通量データや気象データは単一のデータとして分析をしても企業にとって有益な分析結果を得ることが難しいことを示していると言えるだろう．

このように企業間でデータを提供し複合的に分析する取り組みは進んでいないが単一の事業者がさまざまなデータを保有している場合は積極的に活用されているようである．ようである，というのは，このような事業者は，SNS，検索エンジン，携帯電話などサービスのプラットフォームを提供している法人が多く，実際どのように分析し，事業に活かしているか実態を把握するのは難しいためだ．しかし，確実に利用している，と考えられる事例はある．例えば Google では，“Google AdWords[1]” という広告出稿のためのサービスを展開しているが，広告効果の表示画面にクロスデバイスコンバージョンの結果を表示する機能がある．これには，ウェブサイトでの行動，電話での問い合わせ，実店舗への来店，アプリ内の操作，が含まれると説明されている [2]．つまり，Google はさまざまなデバイスで展開される自社のプラットフォーム上で行動するユーザ行動を単一のユーザアカウントに紐付けて追跡し，分析や自社のサービスとして利用していることを示しているだろう．このように，ユーザ行動をあらゆる場所，時間において記録していくことの重要性は増しているが，現状では巨大プラットフォーム事業者が独占している状況である．

[1]Google AdWords: https://www.google.co.jp/adwords/
[2]https://support.google.com/adwords/answer/3419678

1.1.3　複数事業者によるデータの連携と活用

ユーザの行動に基づくデータはさまざまな公的機関，事業者が保有している．しかしながら，各機関，事業者が保有するデータの連携を通じた知識発見は進んでいない．この原因は，本論文では，データの統合・活用に関する技術的な課題及びデータ保有者間の協力に関する障壁が存在することによって，複数のデータを組み合わせ分析することで得られる知見の価値が，データを提供するリスク，コストと比較し，低くなっているためだと考える．本項ではこれらについて説明し，本研究における課題を明らかにする．

(1)　複数のデータを組み合わせた分析の効果

複数データを組み合わせて分析することの価値は事業者にどのように捉えられているのだろうか．図 1.2 にデータ分析を行った企業のうち分析の効果があったと感じた企業となかったと感じた企業をデータの組み合わせ数に応じて比較したグラフを示す．これによると 1 種類のデータのみで分析を行った場合は，分析の効果がなかったと感じる企業が多いが，２種類以上を組み合わせて分析した企業では，効果があったと感じる企業の方が多いことを示している．これは単一のデータによる分析では効果的な分析が難しいことを示す一つの例と言って良いだろう．そして，２種類以上のデータを組み合わせて分析することで抽出される知見は単一のデータで分析した場合と比較し，価値のある知識を抽出する可能性が高い．つまり，これまでよりもっと多くの種類のデータを組み合わせて分析することを促進することで，分析の価値を高めることができる．

(2)　データを提供するリスクとコスト

一方で，事業者は，データを提供するリスク，コストは非常に高いと感じていると考えられる．本項では，プライバシー保護に関するリスクとデータソース間でのデータの紐付けに関するコストについて議論する．

まず，データ提供に伴うプライバシーのリスクについて考察する．データを他事業者に提供する事業者の立場から考えてみたい．まず大規模データの提供にはその所有者にとってリスクがある．具体的な例では，2013 年 07 月に明らかになった JR 東日本が日立製作所に Suica 利用データを販売しようとし，中止となった事例が挙げられるだろう [3]．このように現状では社会的需要があってもパーソナルデータ [4] を

[3] ニュース記事の一例:http://business.nikkeibp.co.jp/article/opinion/20140718/268916/?rt=nocnt

[4] ここで，パーソナルデータとは，ある特定の個人に関連する情報やデータのうち，それ自体は単体では個人の特定，識別に繋がらず，かつ，個人を識別するための情報に紐付けられていないものとする．

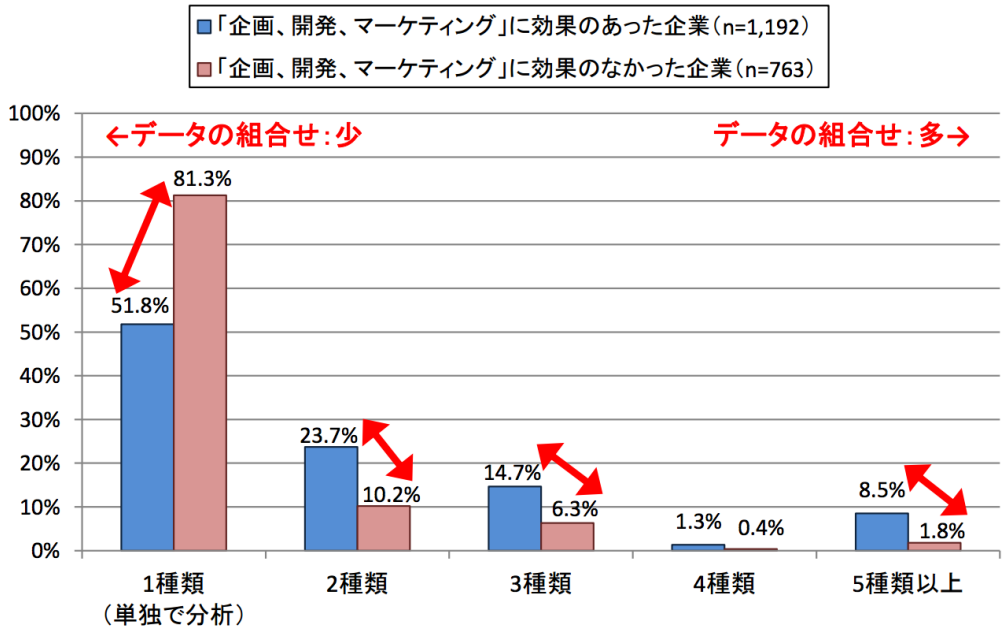


図 1.2: データの組み合わせ数と効果を感じる割合 (「ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究 (2015)」から引用 [85]).

事業として販売することに，社会的合意を得ることは難しい．事業者は，ユーザと Suica データ利用に関して規約を順守した事業展開と考えていたが，ユーザへ説明不足であったと取りまとめている [5]．このような事例もあってデータの所有者は，簡単にはわからないリスクを恐れてデータを提供することをためらっている可能性がある．

次に，データの紐づけの難しさが公的機関，事業者間のデータ連携を困難にしている．各々の公的機関，事業者が蓄積したデータはそもそも他のデータと組み合わせて分析されるために適した形式で蓄積されていない．例えば異なる企業の SNS データと検索ログデータを組み合わせた分析を行う場合には，それぞれのユーザが ID によって対応付けられておらず，分析が困難であることが考えられる．

(3) データ連携と活用の促進のために

ここまであげた課題はユーザを単位としてデータを結び付けることに起因していると考えられる．まず，複数のデータを組み合わせることで，分析結果の価値を向上させることは示した．一方で，リスクとコストの問題のために，事業者間のデータ連携は進まない．それは，Suica の例においてはユーザ単位で分析を行うことがプ

[5] http://www.jreast.co.jp/information/aas/20151126_torimatome.pdf

ライバシーの懸念の根幹にあり，SNS や検索ログデータの連携の例ではユーザの紐づけの難しさが原因である．この課題を解決する方法の一つは，ユーザを軸にデータを分析せず，複数のデータソースの要素の関係性を分析することである．例えば，鉄道の移動データであれば駅を要素とし，SNS や検索データであれば SNS や検索サイト自体を要素とする．これらのデータの構造は大きく異なるが，両社ともユーザの移動（遷移）を介して要素間の関係性を推定することが可能である．しかし，第 2 章で述べるようにユーザ移動（遷移）による要素の抽象化を介して知識を発見する手法の開発は進んでいない．

1.2 本研究の目的

本研究では，これまで述べたように，さまざまな事業者にデータが蓄積されている中で，そのデータの利活用が進まない現状を課題とする．そして，この課題は，複数公共データの統合・活用に関する技術的な課題及びデータ保有者間の協力に関する障壁が存在していることにより，複数のデータを分析することで得られる知見の価値が，提供するリスクと比較し低くなっている，という仮説を設定する．そこで，本研究では，得られる知見の価値を高めるために，増大する異種データを結び付けて知識発見を行う手法の開発と，異種データをもつ公的機関，事業者間が連携をとるために必要なデータ提供のリスクを低減するための方法論を，異種データ間ユーザ移動ネットワークとして提案する．この異種データ間ユーザ移動ネットワークを適用することで，得られる知見が高められることを，異なる 2 つの分析対象 (第 4 章，第 5 章) によって明らかにする．また，データ提供におけるリスク低減に関しては，異なるデータを所有している事業者間の戦略的連携を図る取り組みをもとに，事業者間連携の方法論を第 6 章で提案する．

1.3 異種データ間のユーザ移動ネットワークの提案

本項では，異種データ間のユーザ移動ネットワークという概念について提案する．このネットワークは簡潔にいうと，異なるデータに含まれる要素間をユーザの移動量に応じてネットワーク化する概念である．図 1.3 に，あるユーザが時間経過する中で様々なデータを生み出し，保存されていく様子を模式化したものを示す．これは 1 日の中で，1 人のユーザがチャットツールの 1 つである LINE[6] やマイクロブログサービスの 1 つである Twitter[7] を通して，実世界を移動したり，購買したりする様子を表したものである．LINE を通して誰かとチャットを行えば，それは LINE が

[6] コミュニケーションアプリ LINE:https://line.me/ja/
[7] マイクロブログサービス Twitter: https://twitter.com/

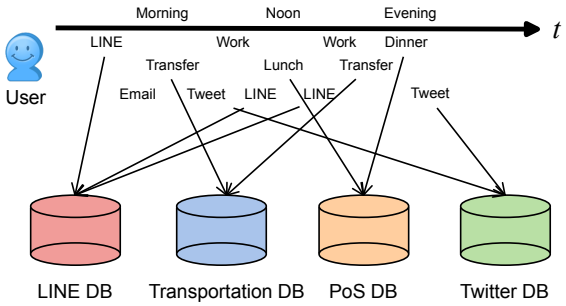


図 1.3: ユーザの日々の行動が様々なＤＢに取り込まれる様子.

管理するデータベースに記録される．また，鉄道による移動を行えば，その移動は利用した鉄道会社の管理するデータベースに記録される．何か商品を購入すればその記録は PoS(Point of Sales) データとして店舗等が管理するデータベースに記録される，といった具合である．このように単一のユーザの全記録に注目し分析していくことを，本稿ではユーザ中心の分析と呼ぶ．

ユーザは 1 日の中の異なる時間に実に多様なデータを生み出している一方で，そのデータを格納するデータベースの種類は生み出すデータの数と比較すると少ない．また，データ保有者は興味対象のユーザが実際に利用しているかどうかよくわからないようなデータを保持している場合も多いだろう．そこで，ユーザ中心の分析に対し，保管されている個々のデータベースを中心とし，その中をユーザが移動するネットワークとして，ユーザ移動ネットワークによる分析を提案する．図 1.4 に異種データ間のユーザ移動ネットワークを示す．この図の中で異種データとは，異なるデータベースに蓄積されたデータのことを指す．各データベースは，データの取得元のデバイスや Web サイトの履歴から取得されたものであり，所有者が異なる．また，各データベース間を結ぶエッジは，ユーザの移動を表している．異種データ間ユーザ移動ネットワークは，図 1.3 のようにユーザを中心にデータ生成を捉えるのではなく，データベースを中心にその中をユーザが移動してデータが生成されていくと考えるものである．例えば，図中の “Data Source A” を “LINE DB”，“Data Source B” を “Transportation DB”，“Data Source C” を “PoS DB” としよう．ユーザを中心とする図 1.4 において，ユーザは朝 LINE でチャットを行ったあと，移動し，昼食をとるといった行動を取り，データを生成している．これをデータベースを中心で捉えるとユーザが LINE DB にアクセスした後，Transportation DB にアクセスし，さらに PoS DB にアクセスしたとみなすことができるだろう．つまり，ユーザ自身が各データベースへ移動していると考えることができる．

このユーザ移動ネットワークを単一のユーザの移動だけでなく大量のユーザの移動を元に作成することも可能であろう．その場合には，データベース間のエッジは，移動量や移動確率，データベース間のつながりの強さを表した指標で接続される．

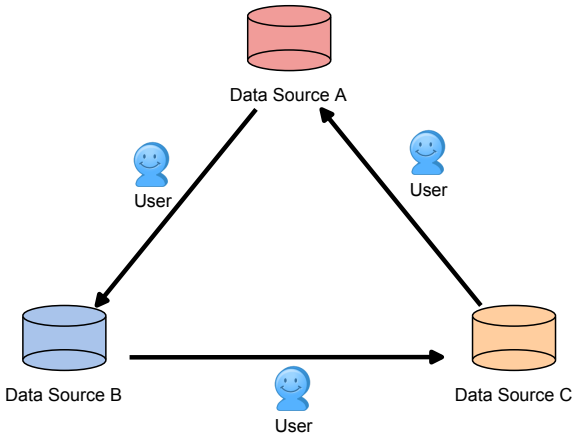


図 1.4: 異種データ間のユーザ移動ネットワーク.

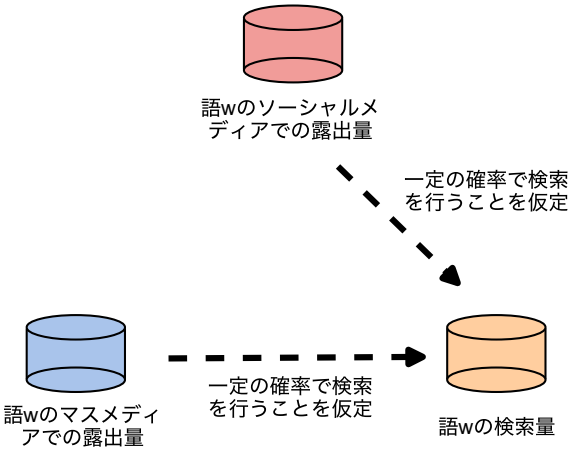


図 1.5: 異種データ間でユーザを部分的に特定できる場合のユーザ移動ネットワークの一例（詳細は第 4 章を参照）.

本研究では，提案した異種データ間ユーザ移動ネットワークを元に，実世界で生成されたデータを元に分析を行い，知識を抽出するタスクを設定した．それぞれのデータベースの保存形式は異なるため，データベース間で同一のユーザの移動履歴を完全に対応させることは困難である．そこで，本研究ではデータベース間でユーザを一意には特定できないが部分的に特定可能な場合と一意に特定できる場合に分け，それぞれについて異種データ間ユーザ移動ネットワークを適用した知識抽出の有効性の検証を行う．ここで部分的に特定可能であるとは，データ間を移動するユーザ群の持つ共通の属性，行動等で集団として特定可能な場合を指す．

まず，ユーザを部分的に特定できる場合のユーザ移動ネットワークの例を図 1.5 に示す．詳細は，第 4 章で述べるが，ここでは複数のメディア露出量と検索量に基づいたユーザ移動ネットワークについて考える．そして，メディア間のユーザ移動に

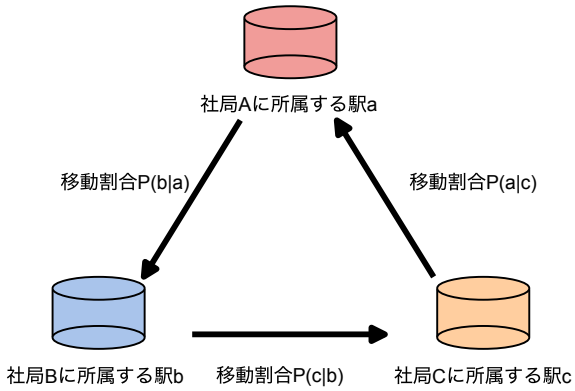


図 1.6: 異種データ間でユーザを一意に特定できる場合のユーザ移動ネットワークの一例（詳細は第 5 章を参照）.

ついて個別には追跡できないことから，あるメディアのユーザは一定の割合で別のメディアへ移動するという仮定に基づいてネットワークを形成している．ここでのユーザはそれぞれのメディアに共通する，ある商品に興味，関心を持つ集団として部分的に特定する．これは例えば，テレビを見て検索をするというような行為が視聴者のうち一定の割合で存在することを仮定している．このようなユーザの移動は商品の特性によって異なってくると考えられる．本章では各ヒット商品におけるユーザの移動ネットワークを推定することで，商品の特性を分類し，どのようにメディアが商品のヒットに影響を与えたかについて分析する．この章では，複数のデータベース間でユーザを追跡することができない場合には，移動に仮説を立て，分析を行うことで知識抽出が行えることを示している．この章で提案した手法は，分析目的に対して保持しているデータが少ない場合でも，それを補うようなデータを追加することでさらに有効に機能する可能性が高いと考えられる．

次に，ユーザを一意に特定できる場合のユーザ移動ネットワークの例を図 1.6 に示す．詳細は，第 5 章で述べるが，ここでは複数の鉄道運営事業者間の各駅間のユーザ乗降ログに基づいたユーザ移動ネットワークについて考える．具体的には，ある駅で乗車したすべてのユーザのうち別のある駅へ移動するユーザの割合でエッジを接続し，ユーザ移動ネットワークを形成している．第 5 章ではこれを元に駅に行く目的を推定するタスクを行っている．この章では，複数のデータベース間でユーザを追跡できる場合には，特に移動の系列の情報（文脈情報）を積極的に利用することを提案している．

異種データ間ユーザ移動ネットワークを適用するそれぞれのケースでわかった内容を簡潔にまとめると以下のとおりである．データ間で一意にユーザを特定できる場合では，データ間のネットワークに仮定を置く必要が無いこと，ユーザ移動の系列情報（文脈情報）を積極的に活用することが重要であること，これにより，各ノー

ドについて分散表現を獲得し，ネットワーク的な近接性をより直感的な記述が可能になることがわかった．一方で，データ間でユーザを一意に特定できず部分的に特定する場合では，データ間の要素の関係性に仮定を置く必要があること，仮定を検証することで，データ間の要素の相互作用の程度を明らかにすることができることがわかった．

最後に，このように提案したユーザ移動ネットワークによる知識抽出の有効性は，実社会でも有用であるか，という議論が残る．そこで本論では，このユーザ移動ネットワークをシステムとして実装し，実務的に日々分析業務を行っている関係者の方を通して有効性の検証を行った．その結果，自らが保有するデータだけでなく他社とのデータを用いることでさまざまな新たな知見を得られること，自社の機密情報の漏洩に対する懸念を解決するためにそれぞれのデータ保有者間で合意を形成し，媒介する調停者を設置することの重要性が明らかになった．

1.4　本研究の貢献

本研究の貢献は以下の３点に要約される．

ユーザを介した複数公共データの統合による知識抽出手法の提案

異種データ間のユーザ移動ネットワークを提案，定義し，複数公共データの統合に適用することで，知識抽出が可能になることを示した．また，それぞれのデータの持つ性質に注目し，データ間でユーザを追跡できる場合とできない場合との２つの場合に関して個別の事例における知識抽出手法について提案を行った．

有効な知識抽出手法の提示

ユーザを追跡できる場合には，文脈情報の利用の有効性を示し，追跡できない場合には興味，行動が共通する集団として特定し，仮定をおいてネットワーク性の検証を行う手法の有効性を示した．

ユーザを介した複数公共データの統合による知識抽出の実社会での応用の可能性の提示

異種データを用いた分析が，実務担当者にとって有益な知見を抽出できることを実例により示し，今後さまざまなデータに対し，『複数公共データの統合による知識抽出』の枠組みが積極的に適用されていく可能性を示した．

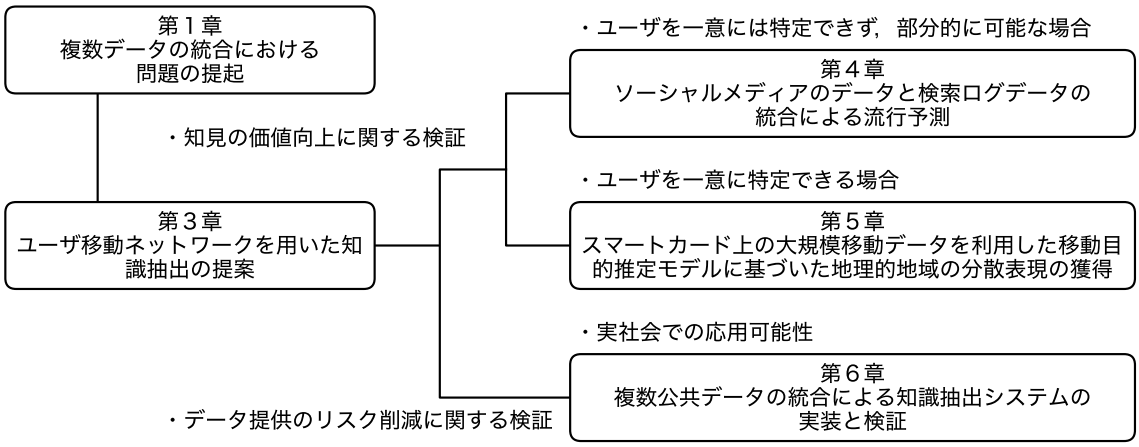


図 1.7: 提案と各章での検証の関係.

1.5　本論文の構成

本論文の構成について述べる．まず，提案と各章で行った検証の関係を図1.7示す．本章では，複数公共データの統合と知識抽出に関する研究の背景と課題について問題提起を行った．特に，公共データの活用が進まない現状に対する課題を，複数公共データの統合・活用に関する技術的な課題及びデータ保有者間の協力に関する障壁の２点にあるということを述べた．次に，第２章では本研究に関連する研究について述べる．第３章ではユーザを介した複数公共データの統合による知識抽出を提案する．特にデータソース間をユーザを介して接続する，ユーザ移動ネットワークという概念を提案し，その適用について議論を行う．そして，第４章ではソーシャルメディアと検索記録の統合について提案した仮説の検証を述べる．この章では，データソース間でユーザを一意には特定できないが，部分的に特定できる場合のユーザ移動ネットワークの適用と知識抽出について説明する．さらに，第５章ではアンケート調査と公共交通データの統合について提案した仮説の検証を述べる．ここでは，データソース間でユーザを一意に特定できる場合のユーザ移動ネットワークの適用と知識抽出について説明する．第６章では提案したユーザを介した複数公共データの統合による知識抽出という概念を具現化したシステムとその検証について述べる．本章では特にデータを組み合わせることで得られる知識の価値，一方でデータ提供のリスクの低減に関して実際の事業者の分析担当者によるシステム利用を元にしたアンケート調査から評価を行っている．その後，個別の章での検証内容から第７章で考察をし，第８章で結論を述べる．

1.6　まとめ

本研究は，データソース間にあるユーザを介したネットワーク性に注目し，ユーザを中心とするのではなく，データソースを中心としたネットワークを利用した知識抽出の有効性を明らかにする．そこで，このような複数の公共データの持つ異種性とそこにあるネットワーク性に着目し，どのような手法等により分析を行い，複数公共データの統合による知識抽出の活動を実社会で有効に機能させるかということを議論する．本研究で提案したユーザを介した複数公共データの統合による知識抽出によって，現在公的機関，事業者が蓄積しているさまざまな大規模データの提供，活用が促進されるものと考えている．そして，複数の公共データを多角的に分析した結果を元に意思決定がされるようになり，より高度かつ効率的なイノベーション計画を策定できるようになると思われる．

第2章　関連研究

本章では，大規模異種データ分析に関する研究の先行研究を整理し，特に本論のテーマであるユーザ移動ネットワークと知識抽出に関する研究の位置づけについて議論する．

2.1　大規模データの分析に関する研究

大規模異種データとはなにか，という議論をする前に，その前提となる大規模データに関する議論から始めたい．大規模データ (Big Data) という用語は，Diebold の調査によれば，もともと学術用語ではなくシリコングラフィクス社のチーフサイエンティストであった J.Mashey が 1990 年代半ばに使いはじめた [16]．ストレージや計算機資源の普及に伴い，蓄積された大規模データそのものに商業的にも，学術的にも大きな注目が集まるようになったのは，2010 年代に入ってからである．特に大規模データ自身が持つ性質，効果について盛んに議論がされるようになった [84, 14, 54]．本稿では，Wu らの提案する HACE 理論 [72] を元に大規模異種データ分析の課題について整理を行う．HACE 理論は，大規模データの持つ特徴を整理し，それぞれとデータマイニングとの間にどのような関係があるかを明らかにしたものである．この理論では，大規模データの持つ特徴を以下の 3 つに整理している．

1. 異種性を持つ大量のデータと様々な次元
2. 分散して自動的に蓄積し，偏在化した制御機構
3. 複雑かつ変化する関係性

第一の “異種性を持つ大量のデータと様々な次元” で用いられている異種性という語は，ここでは異なる情報源から得られるという程度の意味である．特に大規模データは最初から異なる情報源間での分析を想定し，蓄積されているものではないことに注意が必要である．この異種性は，情報源によるデータベーススキーマの違いや対象の表現方法の違いを生む．例えばある人を表現する場合に，アンケートの集計を目的とした情報源の場合には，年齢や性別などデモグラフィックな属性や回答を関係データベースで格納するだろう．しかし，検査や研究の場合には遺伝子配

列の情報で表現したりする可能性もありうる．このように大規模データでは，情報源とその蓄積形式の多様性があり，統合した分析には非常な困難が伴うことを Wu らは指摘している．本論においても，多様な情報源の持つこのような異種性に注目し，統合することでユーザ行動に関する知見をどのように得られるかということを議論する．

2.2 ユーザ移動ネットワークに関する研究

本稿では，複数の情報源間におけるインタラクションを利用するユーザの移動として捉え，それぞれの情報源をノードとするユーザ移動ネットワークとして提案する．この移動は，実際のユーザの物理的な移動だけでなくインターネット上でのサービス間の移動も含む．そこで，本節ではこのユーザ移動ネットワークを 2 つの側面から議論を行う．1 つは情報源間の実際のユーザ移動を分析するという意味でユーザ行動に関する研究から議論を行う．一方で，このユーザの移動は情報源間での情報の伝搬である，と捉えることも可能だろう．そこで，情報伝搬に関する研究からも議論を行う．そしてこの 2 つの側面から本研究の位置づけを明らかにする．

まずユーザ行動に関する研究では，Zafarani らは個人の行動 (Individual Behavior) と集団の行動 (Collective Behavior) という基準で分類している [76]．個人の行動では，個人の行動やコミュニティに所属する可能性を推定するタスク，モデルに関する研究を対象としている．一方で，集団の行動では情報源への移動量や分布を推定するタスク，モデルに関する研究を対象としている．本稿で提案するユーザ移動ネットワークを利用した分析は，ユーザ個人について分析するのではなく情報源間の移動量や分布の分析を行うことを目的としており，ユーザ集団の行動推定に関する研究に位置づけられる．

また，一方で情報伝搬に関する研究は，Guille らが提案している分類では，トピックの検出 (Detecting Interesting Topics)，伝搬過程のモデル化 (Modeling Diffusion Processes)，インフルエンサの同定 (Identifying Influential Spreaders)[21] の 3 つとされている．ここで，トピックの検出は，ある時点でのネットワーク内での話題を分類，抽出することを目的としている．また，伝搬過程のモデル化は，感染症や流行のネットワーク内での拡散過程のモデル化を目的としている．そして，インフルエンサの同定はネットワーク内で影響力を持つノードを同定することを目的としている．本稿で提案するユーザ移動ネットワークをノード間の情報伝搬と捉えることで，情報伝搬の研究に関するこれらの手法はすべて適用可能である．

本稿では，提案したユーザ移動ネットワークを用いて 2 つの分析を行っているのでそれらとの関連を述べる．

まず，第 5 章の「スマートカード上の大規模移動データを利用した移動目的仮説

に基づいた地理的地域の分散表現の獲得」では，地域の果たしている役割について，移動の系列を文脈情報と捉える手法を用いて推定を行っている．これは，情報伝搬の文脈ではトピックの検出に位置づけられる．一般的に，情報伝搬におけるトピック検出は，伝搬する事象の頻度情報を用いる．単一の頻度情報をトピックとして代表させるるもの [59, 45, 61] や，複数の語の頻度情報からモデルを用いてトピックを推定するもの [4, 12] がある．第 5 章で提案する手法では，少量のラベルデータからノード間移動の系列データを文脈情報として利用することで地域の果たしている役割を推定しており，既存研究で一般的に用いられる手法とは異なっている．

次に，第 4 章の「ソーシャルメディアのデータと検索ログデータの統合による流行予測」では，マスメディアとソーシャルメディアのデータを利用して検索量の予測を行っている．この章では，ユーザを集団として捉えメディア間の情報がユーザを介して移動していることを明らかにしている．単一のソーシャルメディア内でのネットワーク構造を利用するだけでは，情報伝搬の説明を完全に行えないことは，Myers らが明らかにしている [49]．彼らは Twitter 上で起こる流行現象の 71%はそのネットワーク内での伝搬として説明できるが，残りはネットワーク外部からの影響で生じることを示した．この章では，ソーシャルメディアだけでなくマスメディアも外因として捉え，検索量の予測を行う．

2.3 異種データ分析に関する研究

異なる種類のデータを組み合わせて，モデルを設計し，分析を行う試みはこれまで数多く行われてきた．ここでは，さまざまな分野で行われてきたこのような異種データ分析に関する研究と本研究の関係について議論する．

まず，情報推薦の分野では，初期の段階ではユーザのつけた商品に対する評価を元に商品間の類似性関連づけることで，高い評価を得られそうな商品を推薦する協調フィルタリングと呼ばれる推薦手法を採用していた [58]．しかし，商品の評価を行う e-commerce(以下，EC) サービス上には，単に評価値だけでなく，それ以外にもさまざまなデータが記録されている．情報推薦では，記録されているデータのうち，単一のものではなく，そうした異なる種類のデータを積極的に推薦の改善に取り入れるようになっていった．例えば商品に対する感想からユーザの嗜好を読み取るような試み [50] や商品に関するさまざまな情報と組み合わせて推薦を行う試み [10] が挙げられる．最近でも Yelp データセット [1] に含まれるレビュー，ユーザの訪問履歴，店舗の位置情報，店舗の分類情報を利用して，次に訪問する店舗を予測する手法が提案されている [69]．

[1]Yelp データセット: https://www.yelp.com/dataset_challenge

次に，ソーシャルネットワークサービス (以下，SNS) を利用する人々をある種のセンサーと捉える概念は，ソーシャルセンサー [57] として知られている．ソーシャルセンサーに関する試みは，SNS 上の記録と実世界上の現象との関連性を分析する研究と捉えることができる．大きく分けて一時的な社会現象との関連性を捉えようとするものと長期間，広範囲の事象の予測に積極的に利用しようとする試みの 2 つに分けられる．一時的な社会現象の検出では，台風 [57] やイベント [70] を捉えようとした研究が挙げられるだろう．長期間，広範囲の事象の予測では，インフルエンザの流行を検索履歴 [20] や Twitter の記録から予測 [13] する取り組みが挙げられる．

一方で，本研究では第 5 章，第 6 章において，公共交通の移動データを用いた行動推定を行っている．交通データを用いたユーザの行動推定に関する研究では，他のデータを組み合わせて推定を行うものも多い．例えば，PoS データと組み合わせて地域を商業地域や居住地域といったものに分類する試み [44]，SNS 上の記録 (FOURSQUARE[2]) と組み合わせて，移動場所を予測する試み [68] 等が挙げられるだろう．

本研究でも交通データという単一の分析ではユーザの移動目的を理解しにくいものに対して，他社が保有する交通データ，路線図のような駅間の隣接情報，そして，パーソントリップ調査に代表される移動の目的を表す情報というような多様な異種データを組み合わせ，行動推定を行っている．この点で異種データを用い実世界上の行動推定を行う研究の 1 つとして位置づけることができるだろう．

2.4 ユーザ行動推定への文脈情報の適用

本研究では，第 5 章において，ユーザ移動ネットワークでの文脈情報の重要性を示している．このような文脈情報の活用は，自然言語処理の分野で発達した手法 (例えば，Mikolov らの Word2Vec という手法が代表的なものとして挙げられる [47]．) をベースにしている．

文脈情報をユーザ行動推定へ活用しようとする取り組みが近年盛んに取り組まれている．単一のデータで Point of Interest(以下，PoI : お店やレストランへの興味を指標化しようとする取り組み) の推薦に適用しようというもの [79, 42, 24]，友人関係と移動データの文脈情報を用いるもの [73] が挙げられる．

このように移動データへの文脈情報の適用は，現在非常に研究分野として注目されており，本研究もこうした取り組みの一つとして捉えることができるだろう．

[2]FOURSQUARE: https://foursquare.com/

2.5　本章のまとめ

本章では，大規模異種データ分析に関する研究の先行研究を整理し，特に本論のテーマであるユーザ移動ネットワークと知識抽出に関する研究の位置づけについて議論を行った．まず大規模データの分析に関しては，HACE 理論を元に大規模データの分析における課題について議論を行った．特に大規模データが持つ異種性と複雑性という特性が分析を困難にしていることを示した．本稿では，この障壁を乗り越えるための一つの枠組みとして,「ユーザ移動ネットワーク」というものを提案する．この概念は既存の研究で明示的に説明されているものは無いが，ユーザの物理的な移動という側面からはユーザ行動推定に関する研究を参考にしていること，また，ユーザを介した情報の伝搬という側面からは，情報伝搬に関する研究と関連があることを述べた．一方で，異なるデータを組み合わせ，新たな知見を得ようという取り組みはネットワーク性に注目した研究に限らず情報推薦やソーシャルセンサーに関する研究においても，盛んに行われていることを示した．本稿ではこのような異種データの持つユーザを介したネットワーク性に注目することで，有用な知見を得られることを示すことが目的である．この枠組みを用いてデータソースを統合することで，行動推定，情報伝搬に関するタスクへの応用が期待できる．

次章で定義するこの「ユーザ移動ネットワーク」に基づいて，以降の章でどのように適用して知見を得るか，ということの議論を行うとともに，この枠組みは，データソースによる企業間提携においても有用であることを示す．つまり，提案する「ユーザ移動ネットワーク」によって，得られる知見の価値をいかに高め，データを提供するリスクを削減し，複数公共データの統合・活用を促進するか，ということについて具体的な議論を行っていく．

第3章　ユーザを介した複数公共データの統合と知識抽出

本章ではユーザを介した複数公共データの統合による知識抽出について述べる．まずユーザが生み出すデータは，様々なデータベースに分散して保存されていることを示し，それらを統合して分析するための枠組みの必要性について述べる．そして，その枠組みの一つとして，ユーザを介した複数データの統合のための異種データ間ユーザ移動ネットワークという概念の提案をする．そして，異種データ間ユーザ移動ネットワークを適用して，いかに知識抽出を行うかの議論を行う．この知識抽出の方法については，データ間のユーザの特定という観点から 2 つの場合にわけて議論を行う．

3.1　異種データ間ユーザ移動ネットワーク

現代において，人々は日々様々なデータを生み出している．例えば，朝目が覚めて，部屋のライトをつければ，部屋で契約している電力メータの消費電力量が変化するだろう．当然この消費電力量は月末の料金請求の元とできるように記録されている．もちろん電力だけでなく，他にもさまざまなデータを人は 1 日の行動の中で生み出している．誰かと電話で通話をすればその通話記録は通信会社に保存され，チャットをすればその記録はサービスの提供会社によって保存され，何かを買えばその購買記録は店舗やクレジットカード会社に保存されているだろう．このように 1 人の

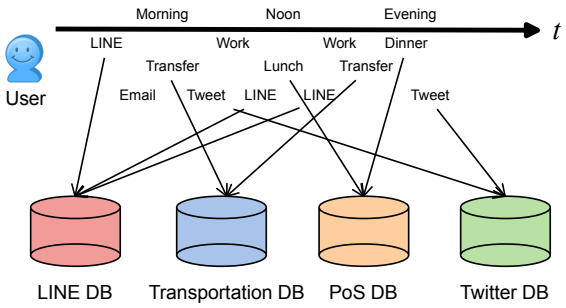


図 3.1: ユーザの日々の行動が様々なデータベースに取り込まれる様子.

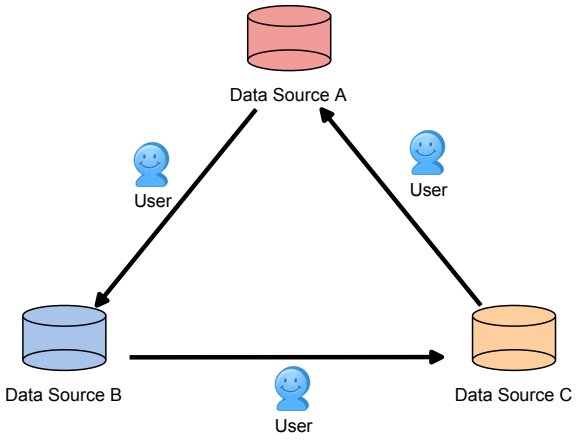


図 3.2: 異種データ間のユーザ移動ネットワーク.

人が時間経過の中で様々なデータを生み出し，保存されていく様子を模式化したものを図 3.1 に示す．これは 1 日の中で，1 人のユーザがチャットツールの 1 つである LINE[1] やマイクロブログサービスの 1 つである Twitter[2]，何らかの公共交通手段を通して，実世界を移動したり，購買したりする様子を表したものである．LINE を通して誰かとチャットを行えば，それは LINE 運営事業者の管理するデータベースに記録される．また，鉄道による移動を行えば，その移動は利用した鉄道会社の管理するデータベースに記録される．何か商品を購入すればその記録は PoS(Point of Sales) データとして店舗等が管理するデータベースに記録される．このようにユーザは 1 日の中の異なる時間に実に多様で大量のデータを生み出している一方で，図に示したようにそのデータを格納するデータベースの種類は生み出すデータの数と比較すると少ない．ユーザの行動分析を行う上で，このようなユーザの生み出すデータをすべて記録することは言うまでもなく有用であろう．また，この図では描けていないデータ（例えば，電力消費量や心拍数などのデータ）もユーザは次々に生み出しているだろう．そうしたデータも含めて分析した方がより詳細な分析が可能になるだろう．しかし，通常それぞれのデータベースの保有者は異なる事業者であることが多い．例えばそれは，交通移動に関するデータを保有している事業者は，PoS や LINE に関するデータは保有していない，ということだ．

このような状況において，果たしてある事業者は他の事業者とデータを相互に提供しあってまで複数のデータを組み合わせて分析を行おうとするだろうか．こうした取り組みは，第 1 章で言及したとおり，現状では少ない．そこには複合的な分析によって，知見を得る枠組みが存在しないことや，実際に得られる知見の価値について不明なこと，さらに個人情報や自社の機密の漏洩に関するリスク，データを提供するための形式の整備に関するコストなど，さまざまな障壁がある．

[1] コミュニケーションアプリ LINE:https://line.me/ja/
[2] マイクロブログサービス Twitter: https://twitter.com/

そこで，本論文では，こうした障害を克服するための 1 つの枠組みとして，データ間に共通するユーザに着目し，ユーザを介した複数データの統合による知識抽出を提案する．具体的には，ユーザを中心として生成されるデータを考えるのではなく，データを記録するデータソースを中心に考える異種データ間ユーザ移動ネットワークという概念を提案する．データソースを中心に考えることで，提供されたデータソース群を元に分析することができる．また，これによって一見関係性やユーザ行動との因果関係が低そうなデータも取り込むことが可能になる．図 3.2 に異種データ間のユーザ移動ネットワークのイメージを示す．この図の中で異種データとは，異なるデータソースに蓄積されたデータのことを指す．各データソースは，データを取得するデバイスの種類であったり，所有者が異なる．そして，各データソース間を結ぶエッジは，ユーザの移動を表している．異種データ間ユーザ移動ネットワークは，図 3.1 のようにユーザを中心にデータ生成を捉えるのではなく，データソースを中心にその中をユーザが移動してデータが生成されていくと考えるものである．例えば，図中の “Data Source A” を “LINE DB”，“Data Source B” を “Transportation DB”，“Data Source C” を “PoS DB” としよう．ユーザを中心とする図 3.2 において，ユーザは朝 LINE でチャットを行ったあと，移動し，昼食をとるといった行動を取り，データを生成している．これをデータソースを中心で捉えるとユーザが LINE DB にアクセスした後，Transportation DB にアクセスし，さらに PoS DB にアクセスしたとみなすことができるだろう．つまり，ユーザ自身が各データベースに移動していると考えることができる．

このユーザ移動ネットワークを単一のユーザの移動だけでなく大量のユーザの移動を元に作成することも可能であろう．その場合には，データベース間のエッジは，移動量や移動確率，データベース間のつながりの強さを表した指標で接続される．

また，図 3.1 では，データベースではなく，データソースという言葉を用いている．本論文では，データソースという語はデータ集合の抽象化された表現，という意味で用いる．データソースは，例えば複数のデータベースを統合したビューやクエリの応答としてのデータ集合などが含まれる．本研究では，このようなデータソース間でのユーザ移動ネットワークを提案する．

次に，異種データ間ユーザ移動ネットワークについて定式化を行う．

ユーザ集合 $\boldsymbol{u}$ のあるユーザ $u_i \in \boldsymbol{u}$ が生み出したすべてのデータ系列を $\boldsymbol{D_{u_i}} = \{d_{i,j} | 1 \leq j \leq N(\boldsymbol{D_{u_i}})\}$ とする．ただし，$N(\boldsymbol{D_{u_i}})$ は，$\boldsymbol{D_{u_i}}$ の要素数を表す．また，データを記録しているデータソース集合を $\boldsymbol{S} = \{S_k | 1 \leq k \leq N(\boldsymbol{S})\}$ とする．ここで，データがある単一のデータソースに記録されている，ということは次のように表される．

$$\exists d_{i,j} \in S_k \Rightarrow d_{i,j} \notin \boldsymbol{S_{\overline{k}}} \tag{3.1}$$

ここで，$\boldsymbol{S_{\overline{k}}}$ は，データソース集合 $\boldsymbol{S}$ のうち，データソース S_k を除いた集合を示す．

このとき，あるデータ $d_{i,j}$ が記録されているデータソースを $S_{d_{i,j}}$ とすると，あるユーザ u_i のデータ系列 $\boldsymbol{D_{u_i}}$ の記録されているデータソースの系列は $\boldsymbol{S_{u_i}} = \{S_{d_{i,j}} | 1 \leq j \leq N(\boldsymbol{D_{u_i}})\}$ と表すことができる．ここで，2 つの連続したデータソースの系列の部分列 $(S_{d_{i,j}}, S_{d_{i,j+1}})$ を**データソース間のユーザ移動**と呼ぶ．この 2 つのデータソース間のユーザ移動量を $Q_{S_k \to S_{k'}}$ で表す．この移動量をどのように定義するかは分析内容，対象によって任意に設定するものとする．例えば，単純なユーザの移動量や移動確率，相関係数のようなデータベース間のつながりの強さを表した指標などが挙げられる．以上のことからデータ間ユーザ移動ネットワークを以下のように定義する．

異種データ間ユーザ移動ネットワークの定義

グラフ $G(V, E)$ において，頂点はデータソースの集合 $V = \boldsymbol{S} = \{S_k | 1 \leq k \leq N(\boldsymbol{S})\}$ とし, データソース間のエッジ集合を $E = V \times V$ とする．
このとき，グラフ内のエッジ $\exists\, e_{i,j} \in E$ は 2 つのデータソース間の移動 $e_{i,j} = (S_i, S_j)$ をあらわす．
そして，エッジの重みの行列を $\boldsymbol{W} = \{W_{i,j} | 1 \leq i, j \leq N(\boldsymbol{S})\}$ とした時，エッジ $e_{i,j}$ の重みを $W_{i,j}$ とし，その重みをユーザ移動量 $Q_{S_i \to S_j}$ とする．すなわち $W_{i,j} = Q_{S_i \to S_j}$ とする．
このような重み付き有向グラフを**データ間ユーザ移動ネットワーク**とする．

グラフィカルモデルの一つにエッジを頂点間の条件付き移動確率で表すベイジアンネットワークと呼ばれるノード間の因果関係の推論を行うモデルがある．本章で提案した異種データ間のユーザ移動ネットワークでもエッジの重みを条件付き確率で表した場合，ベイジアンネットワークとして捉えることが可能である．一方で，本章で提案した異種データ間ユーザ移動ネットワークはこのような因果関係を推定することでは目的ではなく，データソース間の関係に介在するユーザに注目して分析することで，複数データを組み合わせ分析を行い，価値の高い知識抽出を行おうとするものである．

3.2 異種データ間ユーザ移動ネットワークが解決する課題

ここでは，異種データ間ユーザ移動ネットワークによって解決しようとする課題について述べる．それぞれの事業者が相互にデータを提供することを促進する上での障害は，第 3.1 節ですでに述べた．その障害のうち，実際に得られる知見の価値については用いるユーザ移動ネットワークの構成方法，実際の分析に依存するだろう．そこでここでは，データを提供することによって発生するリスクやコストに関

する課題について言及する．

そもそも異種データ間ユーザ移動ネットワークは複数のデータソースを元に何か分析を行う場合に用いられることを想定している．個々のデータソースの接続には以下にあげるような問題がある．

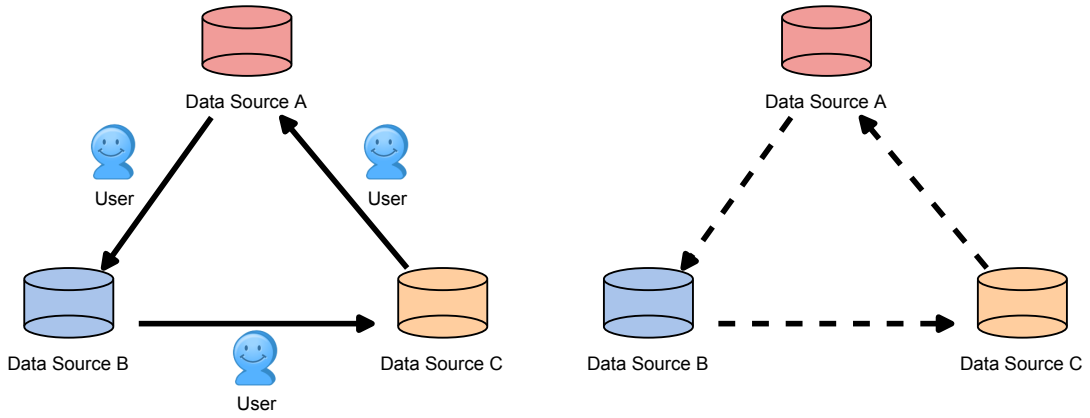
1. データ格納の形式，粒度が異なる．

2. データソースの提供者間に利害関係がある．

3. データソースが不足している，または，無関係のものがある．

それぞれのデータソースは，本論文で想定しているような複合的な解析のために格納されているわけではない．そのデータを取得しているサービスにとって都合の良い形式で格納されている．これは，格納しているデータの内容についても同様である．通常は後のデータ分析のために格納されることは少なく，基本は日常的な運用のための資料として格納されていることが多い．当然それぞれのデータソースはその管理者のものなので，物理的に移動させたり，データソース全体を他のデータソースと同一のデータソースとして統合して扱うことは難しい．さらに，それぞれのデータソースを管理する人または法人が異なることも多い．このような場合には複数のデータソース間に利害関係や機密事項の漏洩リスク，社会的な問題が発生する場合がある．そのため，ユーザ移動の紐付け方法や各々のデータの公開範囲について調整を行う必要がある．これら２つの課題解決をシステムとして開発し，検証を第６章で行った．

データソース間の形式，粒度の問題，利害関係の調整ができたとしても，まだ課題が残っている．それは分析に必要なデータソースが不足している場合である．この場合，ユーザの移動として関係のありそうなデータソースを追加することが想定される．もちろん追加するデータソースは，実際にはユーザの移動とは関係が無い可能性もある．そして，このような追加的なデータソースはデータソース間を移動しているユーザを特定できない可能性がある．完全にユーザが複数のデータソース間で特定できる場合と一部特定可能な場合にわけ，ユーザ移動ネットワークの効果について第４章，第５章で検証を行っている．これらの章では，個別の事例に対し，ユーザ移動ネットワークを適切に適用することによって，高い価値の知識抽出が可能なことを示している．



(a) User Tracked Network

(b) User Untracked Network

図 3.3: 異種データ間ユーザ移動ネットワークの分類．(a) ユーザを一意に特定できる場合のネットワーク．(b) ユーザを部分的に特定可能なネットワーク．

3.3 異種データ間のユーザ移動ネットワークに基づく知識抽出

異種データを組み合わせて分析を行う際，高い価値の知識抽出を行うために，提案したユーザ移動ネットワークをどう適用するかということについて述べる．特に複数の異なる形式で保存されているデータの場合，それぞれのデータ間でのユーザの特定には困難が伴うものも多いだろう．そこで本論では，以下の 2 つの場合で異種データ間のユーザ移動ネットワークの効果の検証を行う．

1. 異種データ間でユーザを一意に特定できる場合
2. 異種データ間でユーザを一意には特定できず，部分的に特定可能な場合

それぞれの場合のユーザ移動ネットワークを図 3.3 に示す．ここで，実線で表されるデータソース間のエッジはデータソース間でユーザを一意に特定できることを示す．一方で，破線で表されるデータソース間のエッジはデータソース間でユーザを一意に特定することはできないが部分的に特定可能な場合のデータソース間の関係性を何らかの形式で表すものである．本稿では，これらのユーザ移動ネットワークを知識抽出にどう適用するかは，個別の事例に基づいて議論を行う．

まず，ユーザを一意に特定できる場合のユーザ移動ネットワークの例を図 3.5 に示す．詳細は，第 5 章で述べるが，ここでは複数の鉄道運営事業者間の各駅間のユーザ乗降ログに基づいたユーザ移動ネットワークを形成している．具体的には，ある駅で乗車したすべてのユーザのうち別のある駅へ移動するユーザの割合でエッジを接続し，ユーザ移動ネットワークを形成している．第 5 章ではこれを元に駅に行く

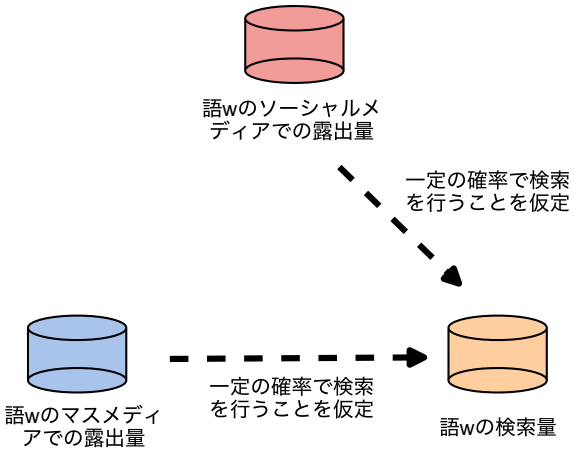


図 3.4: 異種データ間でユーザを一意に特定できないが，一部特定できる場合のユーザ移動ネットワークの一例（詳細は第 4 章を参照）.

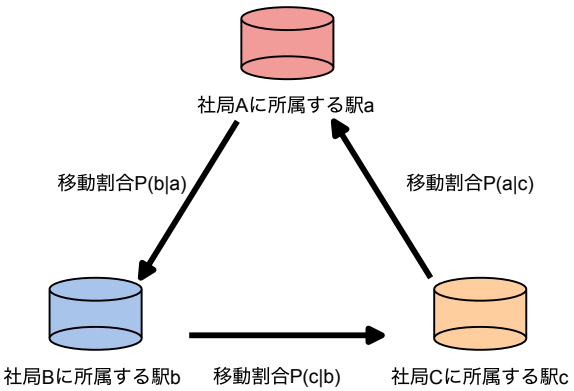


図 3.5: 異種データ間でユーザを一意に特定できる場合のユーザ移動ネットワークの一例（詳細は第 5 章を参照）.

目的を推定するタスクを行っている．この章では，複数のデータソース間でユーザを一意に特定できる場合には，特に移動の系列の情報（文脈情報）を積極的に利用することを提案している．

次に，ユーザを一意に特定できないが，部分的に特定可能な場合のユーザ移動ネットワークの例を図 3.4 に示す．詳細は，第 4 章で述べるが，ここでは複数のメディア露出量と検索量に基づいたユーザ移動ネットワークを形成している．第 4 章では，メディア間のユーザの移動について個別には特定できない．そこで，メディア間のユーザの移動について，一定の割合で移動するという仮説に基づいてネットワークを形成している．これは例えば，テレビを見て検索をするというような行為がテレビを見た視聴者のうち一定の割合で存在する，ということを仮定している．第 4 章ではこのユーザ移動ネットワークを元に複数のヒットした商品について，それぞれのメディアが与える影響を分析し，商品の持つ性質を分類することを行っている．この章では，複数のデータソース間でユーザを一意には特定できないが，部分的に特定可能な場合には，移動に仮説を立て，分析を行うことが必要であることを提案している．この考え方を用いることで，一見無関係と感じられるデータや，格納の形式，粒度が異なるデータであっても仮説を元に追加しておくことで，仮説の検証を行うことが可能になる．これによって，分析目的に対して手持ちのデータが少ない場合でも，それを補うようなデータを追加する可能性が出てくるだろう．

3.4　本章のまとめ

本章では，複数公共データの統合による知識抽出手について提案した．まず，異種データ間のユーザ移動ネットワークを定義した．このネットワークはユーザを中心に考えるのではなく，データを持つデータソースを中心に捉え，その間をユーザが移動していると考えるネットワークである．この異種データ間ユーザ移動ネットワークによってデータソース間の形式の違い，利害関係，データソースの不足に関する問題を解決できることを示した．一方で，複数公共データの統合による知識抽出において，ユーザ移動ネットワークの適用によってどのように高い価値を持つ知識抽出を行うかという点について議論を行った．特に，実際の知識抽出はデータの形式，内容に依存するため，典型的なユーザ移動ネットワークとしてありうる形式として 2 つの場合を想定した．それは，データソース間でユーザを一意に特定できる場合と一意にはできず，部分的に特定可能な場合である．そこで，ユーザを一意に特定できる場合に作られるユーザ移動ネットワークの一例を示し，そこでは，ユーザの移動の系列情報（文脈情報）を積極的に利用することが有用であることを提案した．次に，ユーザを一意に特定できず部分的に特定できる場合には，データソース間の移動関係に仮説を置いて，ネットワークを形成することが有用であることを

提案した．この仮説を検証することによって，ひとまずデータソースを追加し，分析目的に対して有効かどうかを検証することで，データソースの不足を補う可能性があることを示した．

第4章　ソーシャルメディアのデータと検索ログデータの統合による流行予測

4.1　本章で用いるユーザ移動ネットワーク

第3章では，異種データ間のユーザ移動ネットワークに基づいた複数公共データの統合による知識抽出手法を提案し，定式化を行った．そこでは，ユーザ移動ネットワークの作成にはデータソース間でユーザを一意に特定できる場合と部分的に特定できる場合があり，部分的に特定できる場合には，データソース間の関係性に仮定を置く必要があることを述べた．本章では，データソース間でユーザを一意に特定できないが，部分的に特定できる場合のユーザ移動ネットワークの具体例としての分析を行い，知識を獲得することでデータソース間の関係性を示す．具体的には，ヒット商品の予兆をユーザが検索行動に移る前のメディア上の露出による閲覧を元に予測するというタスクを設定する．本章で用いる異種データ間ユーザ移動ネットワークの詳細な図を図4.1に示す．本章では，図に示すように，ソーシャルメディアやマスメディアでの露出量データと，検索量のデータを用いている．そしてこれらのデータ間で同一のユーザを特定することはできない．しかし，これらのデータのうちある商品に関する情報だけを選択し，その商品に関心を持つユーザ群として特定することはできる．そして，ソーシャルメディア，マスメディアのデータと検索量のデータに時間的な差をつけて関係性を調べることで，露出した情報に接触したユーザが検索行動に移っているか調べることができる．つまり，ここではメディア上の露出のあとに検索行動を行うという因果関係の仮定を置いている．本章での分析は，この仮定をメディア間の関係性の程度を示すことで検証することを目的としている．

4.2　本章の背景，目的

近年，ソーシャルメディア，マスメディアが人々に及ぼす影響について盛んに分析されている．その背景の一部として，広報や広告の実務において，商品やサービ

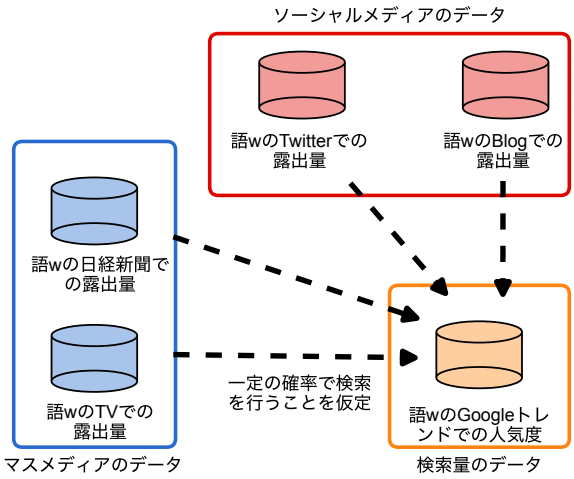


図 4.1: 第 4 章で用いる異種データ間ユーザネットワーク.

スを多くの人々に認知・利用してもらうための効果的なコミュニケーション手段が求められているという状況がある．以前は，新聞，雑誌，テレビなどのマスメディアを通じてしか，多くの人々の認知を促し関心を喚起することができなかった．しかし近年はソーシャルメディアの発達にともなって，マスメディアに留まらず様々なメディア，過程を伴って人々の認知が広がり関心を喚起できるようになってきた．また，インターネットの普及によって，商品やサービスが幅広い消費者の間で話題になり，国境を越えて情報が共有されることも増えつつある．これによって人々は単に消費するという意味での消費者ではなく，消費した商品・サービスを他人に対し発信する役割を以前より果たしていると言えるだろう．本章では，消費だけでなく発信を行う人々という意味で以降，対象とする人々を生活者と呼ぶこととする．

最近では，生活者がソーシャルメディアを通して発信した情報が爆発的に拡散し，消費者の商品・サービス利用に影響を及ぼすことが認識されてきた．そして，その過程について，ダイナミックな分析手法が求められるようになっている．しかしこれまで，このような爆発的な情報拡散 (以降，ヒット現象と呼ぶ) の発生過程について経験則で語られることはあっても，データに基づいた検証がなされることは少なかった．一方，情報工学の分野では，ソーシャルメディア上での情報拡散を動的なモデルを用いて分析を行う試みがなされている．本章では，情報拡散モデルと，複数のメディア上で露出した実際のヒット現象のデータを利用することで，これまでより早い段階でその後の情報拡散過程を説明することを目的とする．また,「口コミ指数」を提案することで，定量的にヒット現象を分類し，どのような情報発信手段を用いれば効果的な情報拡散を可能にするかを明らかにする．

本章で行う研究では，商品やサービスに対する生活者の関心が増大していく過程に，マスメディアやソーシャルメディア上での商品・サービス情報の露出が，どの

ような影響を及ぼしているのかを明らかにする．そのために，商品・サービスに対する生活者の関心をあらわす指標として，Google の検索数を用いる．近年の研究成果では，例えば選挙時の有権者の投票行動は，事前の検索結果から一定程度予測可能なことが示されている [1]．そのため，検索数をもって生活者の関心のみならず，実際の商品・サービス利用状況の指標とすることは妥当であると考えられる．

また，商品・サービスに対する生活者の関心が増大していく過程を分析するに際して，感染症の流行過程をモデル化した手法を用いてソーシャルメディアを利用するユーザ数の推移を予測する研究 [27] の考え方を援用する．この研究は，ある一時点までの検索数の時系列データを元に，その後の検索数の変化を予測するものであり，モデル内部の既知の変数を事後の予測に用いているという特徴がある．そのため，精度の高いモデルを構築するためにはヒット現象が起きた後での予測を行う必要がある．それに対して，本章で行う研究では，ヒット現象が起きる前段階で，その予兆がマスメディア，ソーシャルメディア上で表れていると考えることで，より早期の段階でのヒット現象の説明を行うことを目指した．ただし，ヒット商品は，リリースされた商品のほんの一部であるので，ヒット商品と同様の予兆を示していたが，実際にはヒットしなかった商品が多数存在することが考えられる．ヒットの予兆を示した商品がその後どの程度ヒットするかを予測するには，ヒットの予兆を示したがヒットしなかった商品群を含めた上での分析が必要であろう．しかし，ヒットの予兆というものが明らかになっていない段階で，ヒットしなかった商品群を特定することは難しい．本章で行う研究は，このような理由から，ヒットの程度を予測することに主眼を置くのではなく，ヒットした商品群に共通する，ヒットの予兆とは何かを明らかにすることを目的とする．さらに本章で行う研究では商品・サービスの情報拡散について,「口コミ指数」を提案し利用することで，個別の流行現象をマスメディア上の露出によって流行したタイプと，生活者の口コミによって流行したタイプに分類した．これにより，商品やサービスの持つ性質に応じた商品 PR の方策を検討するための指針とすることができるだろう．

本章は以下のように構成される．まず第 4.3 節で関連研究について述べ，本章で行う研究の位置づけや，有用性や新規性について議論を行う．第 4.4 節で提案手法の背景と根拠について議論し，提案手法について説明を行う．その後，第 4.5 節で使用するデータと実験について説明する．第 4.6 節で分析結果を示す．分析結果について第 4.7 節で考察し，本章で行う研究で明らかになったことを述べる．最後に本章で行う研究についてまとめ，貢献と今後の研究方針について述べる．

[1] 2015 年 1 月 9 日 ヤフー株式会社「第 47 回衆院選の議席数予測を振り返る」
http://docs.yahoo.co.jp/info/bigdata/election/2014/03/

4.3 関連研究

ここでは，関連する研究分野として，メディアが情報流通に及ぼす影響に関する研究，インターネットによる情報流通への影響の研究，情報拡散に関する研究について述べたあと，本章で行う研究の位置づけについて明確にする．

4.3.1 インターネットによる情報流通への影響

1990 年代からのインターネットの普及によって，生活者は自身の持つ情報を全世界に発信できるようになった．遠藤はネットメディアの出現によって誕生した社会を「間メディア社会」と述べている [86].「間メディア社会」では，ネットメディアの誕生によって従来のマスメディアが消滅するのではなく，両メディアが並立することで，情報拡散の回路が拡張され，多層化していく [87].

さらに，インターネットによる情報流通への影響として特筆すべき点は，不特定のユーザによってインターネット上に投稿される口コミの発生である．口コミ (Word of Mouth，以下 WoM) とは，会話を通じた人から人への情報の伝達を指す．特にインターネットを通じた口コミは電子的口コミ情報 (electronic Word of Mouth，以下 eWoM) と呼ばれ [31]，近年のインターネット利用機会の増大とともに重視されるようになってきた．

電子的口コミ情報に関する研究は，大別すると 2 つの内容に集約される．1 つはネット上の口コミの内容に関する研究であり，もう 1 つはネット上に口コミを投稿した人物同士のネットワークに関する研究である．

まず，ネット上の口コミの内容に関する研究としては，投稿された口コミの内容を評判に関する情報として捉え，例えば，肯定的であるか否定的であるかというようなセンチメントを検出する研究が挙げられる．Hatzivassiloglou らや Turney らによって行われたセンチメントを検出する研究は，初期に予め設定した “excellent” や “poor” といった種となる単語を元に，何らかの指標に類似した語，共起する語彙を検出し，文章全体が肯定的か否定的かを判定する [23, 65]．また，商品情報のどの内容に興味をもたれているかを検出する研究分野もあり，Ghose らは，Amazon の本のレビューに対し “heplfulness” であると認めたユーザの数を利用することで，読み手（購買を検討するユーザ）にとって有用な情報を提供する試みを行っている [19]. さらに，実際に投稿された内容を細かく分析することで，分析対象とする商品のどの部分に関心が持たれているかを検出する研究もある．こうした分析は，意見要約に関する研究の 1 分野とみなすことができ，多くの研究が蓄積されている [41].

口コミのネットワークに注目した研究は，Lazarsfeld や Katz らによって提唱された 2 ステップフローコミュニケーションモデルに基づいている [37, 28]．このモデル

は，影響力の強い人物，メディアがそのフォロワーに影響を与えるというもので，影響を受けた人物がさらにそのフォロワーに影響を与えるということを繰り返していくという仮説である．後述するネットワーク研究の進展に伴い，このモデルに基づいてネットワーク上で影響力を持つ人物を検出しようという研究が行われ [6]，Web サービスとしても公開されるに至った [2]．口コミのネットワークに関する分析結果を利用した Web 上のサービスは，マーケティング担当者のソーシャルリスニングのツールとして重要性を増しており，近年注目されている [3]．

しかし，これまで述べてきた従来の研究は，マスメディアや口コミ，インターネット上の口コミの影響についてある一時点や一定期間の分析は可能でも，情報が拡散する動的な過程をモデル化し分析するのには限界があった．

4.3.2 情報拡散過程に関する研究

本章で行う研究では，商品・サービスに関する情報がどのように拡散し，ヒット現象を生み出しているのかについて分析を行う．こうした研究は情報拡散に関する研究の 1 つと捉えることができる．ここでは，特にネットワークに関する分野を中心に，情報拡散に関するこれまでの研究と本章で行う研究の関連性について述べる．

情報拡散モデルはさまざまな研究で提案されており，消費者行動論の文脈で，商品の購買行動が増える過程をモデルとして定式化した研究も多い [55, 17]．特にバスモデルとして知られる購買予測モデルでは，耐久消費財などの新商品の購入意欲を，他人にまどわされない購入意欲（Innovation 効果）と他の購買者数からの影響による効果（Imitation 効果）の和で表現できると提案している [17]．しかし，この形式のモデルは人々のもつネットワークについて十分に考慮していない．

近年，情報拡散過程のモデル化に関連して，世界的にネットワーク分野の研究の進展が見られている．もともと同分野はグラフ理論の研究を基盤とし，近年はコンピュータを利用し実データに応用することで, インターネットのみならず生態系や電力供給網など, 様々な分野で進展をみせている．特に，Watts らによる「スモール・ワールド・ネットワーク」の研究 [71] と，Barabasi らによる「スケール・フリー・ネットワーク」の研究 [3] によって，同分野の研究が一気に進展することとなった．例えば，ツイッターという SNS の持つメディア性，リツイートのネットワークについて議論を行った報告がされている [35]．関連して, ツイッターに注目し，ツイッターを利用しているユーザをさまざまに分類することで属性を明らかにしようとする研究も存在している [2, 26]．

人々の持つネットワークに関するこのような研究の蓄積によって，情報拡散過程

[2] https://klout.com
[3] https://www.hottolink.co.jp/service/kakaricho

のモデル化も進展することになった．SIR モデル [30] は，感染症の流行過程をモデル化したもので，情報拡散過程のモデル化にも応用されている．例えば，John らの研究 [27] では，SNS の利用者数の変化を，『Google トレンド [4]』を元に取得した人々の検索数の履歴から，SIR モデルを拡張したモデルを利用して予測を行っている．また，Takayasu らは，東日本大震災後にネット上で拡散した情報について観察し，報告を行っている [62]．このように SIR モデルは免疫を獲得することで二度と同じ感染症に罹患しないことが前提となっており，これらの分析も一度だけ発生する現象に対して適用されている．

感染モデルに関する調査文献 [9] によると，SIR モデルを始めとした感染症モデルにはさまざまなものがあり，免疫を獲得しない病気に適用可能な SIS モデルや潜伏期間を考慮する SEIR モデルなどが代表的なものとして紹介されている．しかし，本章で行う研究では個々の流行現象を 4.4.5 節において観察した結果，全く同一の内容での流行は起こっていないことがわかった．そのため，同一の流行 (一種の病気とする) には二度 とかからない仮定を置くこととし，SIR モデルを適用することとした．

4.3.3 実際の現象と検索行動の相関

本章で行う研究では，商品に関する生活者の持つ関心の変化のデータとして「Google トレンド」の人気度の時系列の推移を利用している．ここでは，こうしたインターネット上の人々の検索行動と実世界上の人々の行動に相関があることを示す．

代表的な研究として，Ginsberg らが報告している [20] インフルエンザの流行過程に関する研究 (以下，GFT) があげられる．この研究では，実際のインフルエンザの流行と検索数の変化は一致していることが示されている．ただし，この報告では，病院を訪問したインフルエンザの患者数の推移と一致させるために自動的に抽出された複数の検索クエリの推移を多変量解析の手法を用いて最適化させている．これによって，人々の検索行動が実際のインフルエンザの発病と相関していることが認められる．一方で，この手法の精度には Lazer らによって疑問が提示されている [38]．これによると，前の年に作成されたモデルでは，その翌年は高い精度の予測ができなかったことが示されている．これはモデルフィッティングの段階で高い精度を達成するために過剰適合が起こっていることや，メディアでの過剰な露出 (ニュースでの報道) と実際の発症数との乖離が起こっていることが原因である．このように人々の検索行動と実際の行動の相関で汎用的なモデルを構築することは未だ困難を伴うが，少なくともある時点での相関を説明することはかなり高い精度で達成できることが示されている．

今回，我々は商品名を表す単一の検索語のみの「Google トレンド」の人気度の推

[4] https://www.google.co.jp/trends/

移を人々の実際の人気を表すものとして仮定している．GFT では複数の検索語の組み合わせでフィッティングさせることで高い精度を達成しているため，単一の検索語では精度の低下が懸念される．しかし，Mohebi らによって報告された，インフルエンザ以外も含めたさまざまな現象と実際の検索語群の推移との多変量解析の結果を示す「Google Correlate[5]」サービスによる分析結果によると，インフルエンザの場合，“influenza type a”，“symptoms of flu” 等のかなり直接的な検索クエリが 0.9 の相関係数を示していることが報告されている [48]．よって，商品名を表す単一の検索語の「Google トレンド」の結果は，実際の人々の持つ関心と高い相関を持っていると考えられる．

本章で行う研究では，感染モデルと情報拡散に関する研究で得られたモデル構築手法を元に，国内のヒット現象に関する複数のメディア上に露出しているデータを用いることで，ヒット前の段階の様々なメディアへの露出が，人々の関心にどの程度影響を与えるか，または，ヒット現象がどのように推移するか説明できることを示す．

4.4 提案手法

本章で行う研究では,『日経トレンディ』が発表した “2014 年ヒット商品ベスト 30[6]” の上位 10 商品を取り上げ，ヒット前のメディアの露出の中から，商品のヒット現象の過程に早い段階で影響を及ぼす指標を明らかにする．以下では，そうした分析を可能とする原理について説明を行い，指標を提案する．

4.4.1 手法の動機

商品・サービスに関する情報の拡散過程に関するモデルとして，バスモデル [17] や SIR モデル [20] が挙げられる．これらのモデルは，商品・サービスの購入には，ユーザ間でのコミュニケーションが影響を及ぼしているという仮定を伴っている．石井ら [90] はユーザ間のコミュニケーションの効果を 2 つに分けて，バスモデルを拡張している．1 つはユーザ同士の直接の口コミ等の影響である「直接コミュニケーション」，もう 1 つは，ユーザ同士が様々なメディアから受ける影響である「間接コミュニケーション」である．間接コミュニケーションは，対象への評判やうわさを表し，ソーシャルメディア上の口コミの効果を包含している．彼らはこの「間接

[5]ユーザが時系列データをアップロードすると検索クエリ群の時系列データとの多変量解析を行った結果を提示するウェブサービス．それぞれの検索語と入力データとの相関係数を確認できる．https://www.google.co.jp/trends/correlate

[6]http://trendy.nikkeibp.co.jp/article/pickup/20141030/1061085/?rt=nocnt

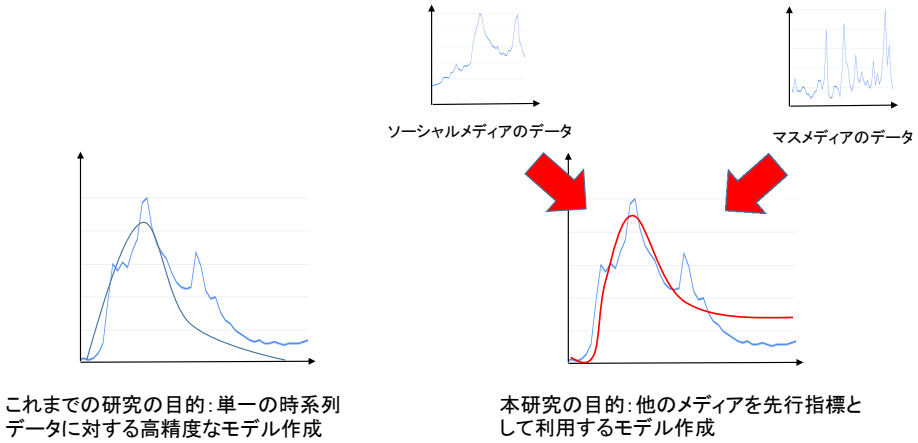


図 4.2: 本章で行う研究の目的と過去の研究との違い．

コミュニケーション」の重要性を主張しており，この効果を表す項を追加することによって，売上の推移データを高い精度でモデル化できることを示した．

しかし，我々は予測したいデータのみを用いて必ずしもモデリングを行う必要は無い．例えばある映画の入場者数を前もって予測したいと考えた場合，ツイッターやブログでの言及，TV での露出量は先行指標として有用であろう．特にソーシャルメディア上では公開前にファンの人々のコミュニティで盛り上がっているだろうことは想像に固くない．そこで，本章で行う研究では，こうしたマスメディアやソーシャルメディアでの露出量を用いて，予測したい対象への情報拡散現象の発生をモデル化することを目指す．本章で行う研究の目的と過去の研究の目的との違いを図 4.2 に示す．我々は図に示すようにソーシャルメディアやマスメディアでの露出を先行指標として利用することで，対象とするデータの推移の説明を可能にするモデルの作成を目指す．

説明したい対象は購買数や売上高の推移であるが，実際のデータを自由に取得することは難しい．そこで，Kotler らの提唱した 5 段階の購買プロセスのモデル [33] や AIDMA，AISAS などの購買意思決定モデルに基づいて，購買の一段階前のデータを利用する．AISAS モデルは，近年のインターネット環境の普及を前提としたモデルであるので，今回はこのモデル内の「Action(購買)」の前段階である．「Search(検索)」の数を実際のユーザの関心を表す指標と捉え，この推移を説明する指標を提案することを目指す．つまり，特定の商品・サービスについて検索を行った人の数を，当該商品に一定の関心を持った人の数 (と相関関係の強い変数) と想定して議論を進める．

本章で行う研究では，人々への関心の拡散を感染モデルとして捉える．そして，ヒット現象はたくさんの人々が同時に興味・関心を持っている (いわば，ある種の病気に感染している) 状態と考える．感染モデルでは，たくさんの人々が同時に興

味・関心を持っているヒット現象が起こっている状態は，興味を持ちやすく，飽きのこない商品を多くの人々が認知していることで発生していると解釈することができるだろう．つまり,「興味を持つ可能性のある人の総数」,「興味・関心の持ちやすさ」,「飽きやすさ」という 3 つのパラメータによって決定される．そして，本章で行う研究では，これら 3 つのパラメータを決定づける要因が，他のメディア上での情報の露出に表れていると仮説を設定し，早い段階でのメディアの露出から提案する指標とこれらのパラメータとの相関の分析を行う．これによって，より早い段階でのヒット商品・サービスの情報拡散過程を説明できるモデルの構築を図る．

なお，感染現象を説明するモデルには，SIS モデルという同じ病に繰り返し罹患するモデルもある．しかし，本章で行う研究では SIR モデルを適用した．これは，1 つの内容に関するヒット現象は一度しか起こらないという仮定にもとづいているからである．この理由を以下に説明する．

4.4.2 SIR モデルを適用する理由

例として「妖怪ウォッチ」の Google トレンドにおける人気度の変化を図 4.3 に示す．図 4.3 に示したとおり,「妖怪ウォッチ」は複数の人気度のピークを持っており，長い期間に，繰り返し流行していることがわかる．さて，それぞれの人気度のピークにおいて，人々がどういう興味を持っているのだろうか．簡易的に確認するために，流行時点で人々がどのような検索語を実際に入力したかの調査を，Google トレンドを利用して行った．Google トレンドでは検索語と共起する語を確認することができる．ここでは，そのときどきの人々の関心を調査するために，ピークの日を含む月に検索語の「妖怪ウォッチ」とともにそれ以前とくらべて検索されている語を確認した．その結果を表 4.1 に示す．表 4.1 を見ると，2014 年 5 月では,「ゲラゲラボーのうた」(妖怪ウォッチの主題歌のこと) や「キャラ弁」(コンテンツに登場するキャラクターを模した弁当のこと) について興味を持たれていることがわかる．次に 2014 年 8 月では,「ガブニャン」(ゲーム「妖怪ウォッチ 2 元祖，本家 (2014 年 7 月 10 日発売)」に登場するキャラクターのこと，希少性が高い),「マクドナルド」(当時，マクドナルドの商品が妖怪ウォッチ関連のグッズとタイアップするキャンペーンを行っていた) に対して興味を持たれていることがわかる．2014 年 12 月では,「ジバコマ」,「ダークニャン」(ともに 2014 年 12 月 13 日発売のゲーム「妖怪ウォッチ 2 真打」に登場するキャラクターのこと，ゲーム内で主人公と友人関係を形成する方法が特殊) に対して興味を持たれていることがわかる．このように繰り返し流行している「妖怪ウォッチ」において，そのときどきにおいて人々が関心を持っている内容はすべて異なっている．

この調査を今回対象とする商品群に行った結果，すべての商品の持つそれぞれの

ピークに対して，すべて異なる内容で流行が起こっていることを確認した．この調査結果を元に，本章で行う研究では 1 つの商品に対して繰り返し起こる流行 (人気度のピーク) を 1 つの現象として捉えるのではなく，すべて異なる現象として捉えることとした．つまり，1 つの流行は 1 つの人気度のピークを持つもので，厳密な意味 (商品のある側面に感心を持たれている流行現象) で同一の流行現象は起きないと考え，それぞれの流行に対して，SIR モデルを適用することとした．

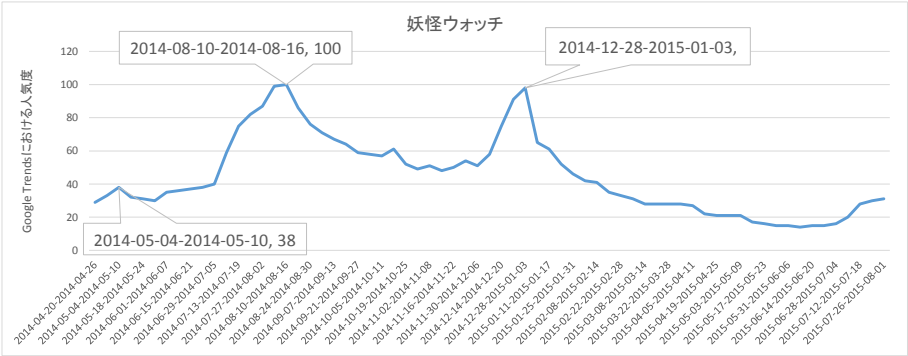


図 4.3: “妖怪ウォッチ” の Google Trends 上の人気度の時系列の変化を表したグラフ．

表 4.1: Google Trends 上で急上昇ワードとして上がっているクエリ．

順位	2014 年 05 月の上昇クエリ	上昇率
1	ゲラゲラポーのうた	+160%
2	ゲラゲラポー	+150%
3	妖怪ウォッチ時計	+80%
4	キャラ弁	+70%
5	妖怪ウォッチキャラクター	+70%

順位	2014 年 08 月の上昇クエリ	上昇率
1	ガブニャン	over +5000%
2	マクドナルド妖怪ウォッチ	+300%
3	マック	+250%
4	妖怪ウォッチラムネ	+250%
5	マック妖怪ウォッチ	+190%

順位	2014 年 12 月の上昇クエリ	上昇率
1	ジバコマ	over +5000%
2	ダークニャン	+600%
3	妖怪ウォッチ真打	+300%
4	真打	+300%
5	真打妖怪ウォッチ	+300%

4.4.3 SIR モデル

本章で行う研究では，SIR モデルによって人々の関心の変化の近似を行う．ここでは，SIR モデルの説明とパラメータの推定方法について説明を行う．

SIR モデルは，Kermack ら [30] によって 1927 年に発表された一階微分方程式によって記述されるモデルで，感染症の流行過程を近似させることを目的としたものである．SIR モデルは以下の 3 つの一階微分方程式によって表現される．

$$\frac{d}{dt}S(t) = -\beta S(t)I(t) \tag{4.1}$$

$$\frac{d}{dt}I(t) = \beta S(t)I(t) - \gamma R(t) \tag{4.2}$$

$$\frac{d}{dt}R(t) = \gamma R(t) \tag{4.3}$$

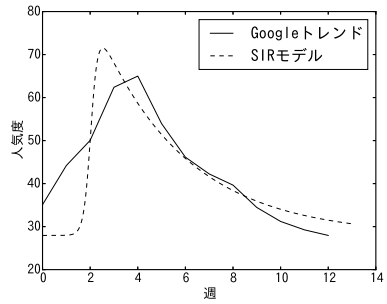


図 4.4: “アナと雪の女王” の人気度と SIR モデルのパラメータを最適化したグラフ.

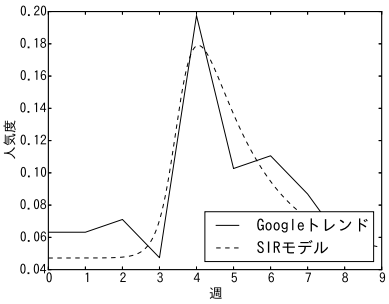


図 4.5: “クロワッサンドーナツ” の人気度と SIR モデルのパラメータを最適化したグラフ.

SIR モデルでは，一定数の感受性保持者 $S(t)$ が感染者 $I(t)$ から一定の割合 β で感染し，感染者 $I(t)$ から一定の割合 γ で回復し，免疫保持者 $R(t)$ となる仮定をモデル化したものである．本章で行う研究では，4.4.5 節の分析の結果，人々は同一商品・サービスについて一度感染した (関心を持ち検索した) 後に，同じ話題で再び感染しない (関心を持たず検索しない) ことが観察されたため，繰り返し感染することの無い SIR モデルを採用している．ここで，各式中の β を感染率，γ を回復率，$S_0 = S(0)$ を初期感受性保持者とする．本章で行う研究では，人々がある商品・サービスの情報に関心を持ち検索した状態をある種の病気に感染している状態とみなす．つまり，ある時点 t で関心を持ち検索を行う人々の数を感染者数 $I(t)$ とする．同様に類推すると，感染率 β は値が大きいほど急激に人々の関心を得ることになるので，情報の拡散力を示す．また，回復率 γ は値が大きいほど急激に人々の関心が無くなることになるので，情報拡散の減衰力を表す．そして，初期感受性保持者数 S_0 はその商品に関心を持ち検索する可能性のある人々の総数を表す．この SIR モデルを利用して，実際の観測データから最小二乗法を利用してパラメータを推定を行う．

本章で行う研究では,「Google トレンド」上でのそれぞれの商品名をクエリとして，ある期間の検索量の推移を取得し，その値を $I(t)$ の観測データとする．このデータを元に，パラメータとして，初期感受性保持者数，感染率 β，回復率 γ の推定を行う．

4.4.4 マスメディア上の露出による情報拡散と生活者の口コミによる情報拡散

ここでは，情報拡散の過程について，企業のニュース発と生活者の口コミ発の情報拡散の 2 種類のタイプに分けられることを，具体的な事例分析をまじえて説明する．

例として，図 4.4，4.5 に「アナと雪の女王」,「クロワッサンドーナツ」の『Google トレンド』上の人気度の変化と SIR モデルのパラメータ最適化を行った結果を示す．「アナと雪の女王」は，2014 年 4 月 6 日～2014 年 7 月 19 日までの期間で,「クロワッサンドーナツ」は，2014 年 10 月 5 日～2015 年 1 月 17 日の期間の検索量の推移を抽出したものである.「アナと雪の女王」に関して，この期間は映画の全国公開から 4 週目にあたり，4 月 26 日からは 3D 日本語吹替版が約 138 劇場で，映画館の観客が一緒に挿入歌を歌える限定公開イベント「“みんなで歌おう♪” 歌詞付き」版が約 85 劇場で上映開始されたことが報道されている．一方,「クロワッサンドーナツ」に関しては，この期間に 2014 年度のヒット商品ランキングが発表されている．SIR モデルによって推定されたパラメータは,「アナと雪の女王」が，$S_0 = 52.1, \beta = 7.05, \gamma = 0.270$ で,「クロワッサンドーナツ」は，$S_0 = 0.231, \beta = 4.56, \gamma = 0.67$ であった．

観測された人気度の変遷をグラフから読み取ると,「アナと雪の女王」は「クロワッサンドーナツ」と比較すると，人気度の総和（各週ごとの人気度の値の和）が高く，変化が緩やかであることが読み取れる．この違いは推定したパラメータに表れており,「アナと雪の女王」は「クロワッサンドーナツ」と比較して，初期感受性保持者の量を表す S_0，拡散の速度を表す β と人気の減衰速度を表す γ が低くなっている．そして，初期感受性保持者の量を表す S_0 は，β が γ に比して十分大きい場合には最終的には，大多数の S_0 が感染する．今回も同様に図のどちらの場合も，推定されたパラメータの β が γ と比較し大きいため，図中の人気度の総和が初期感受性保持者 S_0 の量を概ね比較しているといって良いだろう．よって以降，特別の場合を除き S_0 の量を人気度の総和と呼ぶ．

ところで,「アナと雪の女王」のヒット前のブログの平均露出数は，18,407.0 であったのに対し,「クロワッサンドーナツ」のヒット前のブログの平均露出数は，268.8 であった．このことから，ブログへのたくさんの露出は，S_0 に正の影響を与えることが考えられるだろう．また，それぞれのブログでその期間に見られるようになった語を調査してみると表 4.2,4.3 のようになった．これによると,「アナと雪の女王」では主題歌に対して，好感を示す単語が見られ，イベントが成功していることが観察されるのに対し,「クロワッサンドーナツ」ではミスタードーナツでクロワッサン生地使用している他の商品が多く観察される．この違いは,「クロワッサンドーナツ」が，マスメディアや広告の露出が元で話題になっている可能性が考えられるだろう．そこで，マスメディア上の露出数とソーシャルメディア上での露出数を元に「口コ

ミ指数 (I_{WoM})」を以下の式 4.4〜式 4.5 のように提案する．

表 4.2: Blog 上で「アナと雪の女王」に関連する上昇ワード．

順位	クエリ	上昇率
1	アナ	+1.3%
2	雪の女王	+1.3%
3	主題歌	+1.1%
4	海外	+1.0%
5	アナ雪	+0.8%
6	好き	+0.8%
7	話題	+0.7%
8	子供	+0.5%
9	大ヒット	+0.5%
10	ネット	+0.5%

表 4.3: Blog 上で「クロワッサンドーナツ」に関連する上昇ワード．

順位	クエリ	上昇率
1	ミスド	+2.4%
2	カップケーキ	+1.9%
3	ドーナツ	+0.7%
4	マロン	+0.6%
5	アップル	+0.5%
6	楽しい	+0.5%
7	ミルク	+0.5%
8	うまい	+0.5%
9	美味しい	+0.4%
10	おやつ	+0.4%

$$I_{S_i,M_j} = \frac{C_{S_i}+1}{(C_{M_j}+1)\times R_{S_i,M_j}} \tag{4.4}$$

$$I_{WoM} = \frac{N_S \times N_M}{\sum_i^{N_S}\sum_j^{N_M}\frac{1}{I_{S_i,M_j}}} \tag{4.5}$$

$$I_{blog,MetaTV} = \frac{C_{blog}+1}{(C_{MetaTV}+1)\times R_{blog,MetaTV}} \tag{4.6}$$

$$I_{Twitter,MetaTV} = \frac{C_{Twitter}+1}{(C_{MetaTV}+1)\times R_{Twitter,MetaTV}} \tag{4.7}$$

$$I_{WoM} = \frac{2}{\frac{1}{I_{blog,MetaTV}}+\frac{1}{I_{Twitter,MetaTV}}} \tag{4.8}$$

この式において，S_i は N_S 種類あるソーシャルメディアのうち，i 番目のソーシャルメディア S_i を表し，M_j は N_M 種類あるマスメディアのうち，j 番目のマスメディア M_j を表している．そして，I_{S_i,M_j} はソーシャルメディア S_i とマスメディア M_j 間の露出量 C_{S_i}, C_{M_j} の比率を表している．ただし，式中の R_{S_i,M_j} はソーシャルメディア S_i とマスメディア M_j 間の差を調整する係数で，手動で設定を行う．そして，最終的な口コミ指数 I_{WoM} は，複数のソーシャルメディア，マスメディア間の露出量の比率の調和平均によって算出する．

今回，ソーシャルメディアとして Blog，Twitter，マスメディアとして TV を採用した．特にマスメディアとして TV のみを利用しているのは，調整係数を手動で設定している点と，より多くの人が接触していることを考慮したためである．よって，I_{S_i,M_j},

I_{WoM} は式 4.6–4.8 のようになる．式 4.6，4.7 はそれぞれ Blog と Twitter の露出量と TV での露出量との比である．露出量の尺度は様々なものが考えられ，適切な尺度については議論の余地があるが，今回は単純に出稿数とする．つまり，Blog の場合は該当の記事数，Twitter の場合は該当のツイート数，TV の場合は該当の番組数とした．$R_{blog,MetaTV}$,$R_{Twitter,MetaTV}$ は，TV での露出数がソーシャルメディア上での露出数と比較すると極端に少ないため，調整するために $R_{blog,MetaTV} = 750$,$R_{Twitter,MetaTV} = 950$ を代入している．また最終的に算出する I_{WoM} は，$I_{blog,MetaTV}$ と $I_{Twitter,MetaTV}$ の調和平均を取った値となっている．これによって極端に Twitter と Blog で異なった値となっていても小さい値の方に近い値となる．この口コミ指数が小さい値の場合，その流行に関して，Blog でも Twitter でも TV での露出量と比較して，あまり話題になっていないことを示す．逆に大きい値の場合，その流行は，Blog でも Twitter でも TV での露出量と比較して，大きな話題になっていることを示す．

4.4.5 商品情報の拡散過程分析

前述のように，ヒット商品・サービスの中で，同一の情報発信内容についてとりあげられた期間のデータのみを抽出し，『Google トレンド』のデータを人々の関心の広がりを示すデータとして用いて，SIR モデルのパラメータの推定を行う．次に，検索数が最も多くなったピークの時点（これを流行が頂点に達した状態とする）より以前の一定期間について，各メディアでのその商品・サービスの情報の露出数のデータを取り出し，指標化する．ここでは，以下の指標を定義する．

$$\overline{C} \;=\; \sum_{t=1}^{N} \frac{C(t)}{N} \tag{4.9}$$

式 4.9 は人々の関心が最も高くなった時点までの前 N 時点でのあるメディアでの露出量 $C(t)$ の平均を示したものである．

最後に提案した指標と，商品情報の拡散過程を表すパラメータとの相関を示すことで，類型化した商品・サービスに関する情報の拡散過程を説明する指標を明らかにする．相関係数 r にはピアソンの積率相関係数を採用し，以下のように算出する．

$$r(\boldsymbol{x},\boldsymbol{y}) = \frac{\sum(x_i-\bar{\boldsymbol{x}})(y_i-\bar{\boldsymbol{y}})}{\sqrt{\sum(x_i-\bar{\boldsymbol{x}})^2}\sqrt{\sum(y_i-\bar{\boldsymbol{y}})^2}} \tag{4.10}$$

この式内で，$\boldsymbol{x}$,$\boldsymbol{y}$ はそれぞれ指標群や推定したパラメータ群を表し，x_i, y_i は抽出した期間での指標，推定したパラメータの値を示すものとする．$\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}$ はそれぞれの指標，パラメータの期間全体での平均を表すものとする．

具体的には，4.6.1 節に示す手法で同一の話題となっている期間を同定し，その期間内でパラメータ (S_0,β,γ) の推定を行い，それぞれ $\boldsymbol{x}$ の各要素 x_i とする．次に式 4.9 に示したあるメディア (ブログ，ツイッター，新聞，TV) でのヒット前露出量に関する，期間 i における指標の値を y_i とする．そしてパラメータ，メディアごとの組で，全体の相関係数 r を式 4.10 によって算出する．

4.5 使用するデータ，実験設定

本章で行う研究の分析で使用するデータ，実験設定について説明する．

ヒット商品・サービスについては，“『日経トレンディ』2014 年ヒット商品ベスト 30” の上位 10 商品を選定した．日経トレンディは日経 BP 社が発刊する月刊誌で，毎年 12 月号にその年にトレンドになったものを選出する “ベスト 30” を掲載している．2014 年度のトップ 10 の商品・サービスとその説明を表 4.4 に示す．

本章で行う研究では，さまざまなサービスからデータを取得しているが，すべてのデータを，2010 年 07 月 04 日から 2015 年 08 月 01 日の期間で取得し，1 週間ごとに集計したものを使用する．そして，集計したデータを時系列データとして取り扱っている．また，データ取得には各サービスにおいて商品名を直接のクエリとして入力し，出力されたデータを取得している．

本章で行う研究では，商品･サービスへの人々の関心をあらわす指標として,『Google トレンド』のデータを使用した．ただし，05 位の “Ban 汗ブロックロールオン” は十分な検索量がなかったため『Google トレンド』でデータを取得できなかった．このため，分析対象から外すこととした.『Google トレンド』では，検索量を絶対値ではなく，相対的な値として，0〜100 の範囲で表示を行う．本章では，２つの商品をこの『Google トレンド』上で分析を行うことで，どちらの検索量が多いかの比較を行う．特に対象の期間の中での最大値の比較を行うことで，相対的な検索量の算出を行うこととした．期間中,「妖怪ウォッチ」に関する検索量が最大であったので，この最大値を基準 (1.0) として，他の商品・サービスの検索量の期間中の最大値の算出を行った．この値を比較値（期間中の人気度の最大値の「妖怪ウォッチ」の人気度の最大値を基準とした相対的な値）として，表 4.4 に示す．この値はそれぞれの商品・サービスの期間中の検索量の最大値で，本章では，この値をある時点での人気度（「妖怪ウォッチ」との比較値）として扱う．本章で行う研究で使用する検索量データは，それぞれの商品・サービスの単一での『Google トレンド』での結果 (0〜100) を 0〜1 に正規化を行い，その後この比較値を乗ずることで，各時点での人気度の算出を行っている．

さらに，人々の検索量（人気度）の変化を説明できる可能性のある変数として，マスメディア，ソーシャルメディアにおける露出数のデータを利用する．ソーシャ

ルメディアとして，ブログ，ツイッターを選定し，これらの過去のデータを利用する．ブログは，個人または団体が比較的長い文章を書くもので，本章で行う研究ではAmeba ブログ[7] や livedoor ブログ[8] のデータを利用している．ツイッターは，個人または団体が 140 字以内の短い文章を書くもので，本章で行う研究では 10%サンプリングしたデータを利用している．今回ブログ，ツイッターのデータは株式会社 ホットリンクが運営している “口コミ係長” というサービスから提供されたデータを用いている．マスメディアとしては，新聞，テレビを選定し，これらの過去のデータを利用した．例えば，日経新聞の露出データを日経テレコン[9] から取得した．なお，『日経』を選択したのは，経済メディアが商品・サービスのヒット現象に及ぼす影響を分析するためである．テレビの露出データは株式会社エム・データが提供しているテレビ情報検索システム MetaTV[10] から取得したデータを使用した．対象は在京の NHK 及び民放の全ての番組である．ただし，MetaTV のデータには番組での露出情報は含まれるが CM の放映や視聴率は含まれていない．そのため，本章でのマスメディアの露出に関する分析では CM や視聴者数に関するものは行っていない．

4.6 分析結果

ここでは，提案手法を取得したデータに用いて行った分析結果について議論する．

4.6.1 同一の話題となっている期間の抽出

まず，収集したデータを用いて，同一の話題となっている期間の抽出を行った．これは Google トレンドの推移でピークとなっている週を抽出し，その週の 4 週間前から 10 週間後までの，合計 15 週間を 1 つの流行の期間とした．そして，それぞれの話題について，各メディアの露出状況を調査し，4.4.4 節で提案した「口コミ指数 (I_{WoM})」を算出し，主にマスメディア上の露出によって流行しているものと生活者の口コミによって流行しているものに分類した．ここでは，$I_{WoM} \geq 1.5$ なるものを生活者の口コミによって流行したもの，$I_{WoM} \leq 0.5$ となるものをマスメディア上の露出によって流行したものと分類した．

表 4.5，4.6 に結果の一部を示す．まず，全ての商品・サービスにおいて，生活者による検索数は時間と共に変化していることが確認された．抽出した期間におけるメディアそれぞれに対して，SIR モデルの最適化によるパラメータ推定を行い，後

[7] http://ameblo.jp/
[8] http://blog.livedoor.com/
[9] https://t21.nikkei.co.jp/g3/CMN0F11.do
[10] http://mdata.tv/metatv/

表 4.4: 日経トレンディ2014 年ヒット商品ベスト 30 の上位 10 商品.（ただし，表内の WWoHP は，ウィザーディング・ワールド・オブ・ハリー・ポッターを表す.）

順位	比較値	商品	説明
01 位	6.50×10^{-1}	アナと雪の女王	ウォルト・ディズニー・アニメーション・スタジオ制作の映画.
02 位	1.000	妖怪ウォッチ	レベルファイブが開発したゲームおよび様々なメディアで展開される一連の作品.
03 位	7.80×10^{-2}	WWoHP	テーマパーク施設のユニバーサルスタジオジャパンにあるアトラクション.
04 位	6.24×10^{-3}	ジェルボール洗剤	ゼリー状の球形の新型洗剤.
05 位	—	Ban 汗ブロック ロールオン	直塗りタイプの制汗剤．ボール付き容器が特徴.
06 位	2.43×10^{-3}	伊右衛門 特茶	“特定保健用食品” の許可証票の表示許可を受けた飲用茶.
07 位	4.49×10^{-1}	TSUM TSUM	LINE 株式会社が公開しているスマートフォン向けアプリケーション.
08 位	7.90×10^{-3}	クロワッサンドーナツ	ニューヨーク発祥のクロワッサン生地のドーナツ.
09 位	1.88×10^{-2}	格安スマホ	MVNO（仮想移動体通信事業者）が提供する通信サービスを利用するスマートフォン.
10 位	5.53×10^{-2}	あべのハルカス	大阪市阿倍野区に立地する超高層ビル.

の分析を行った．次に，生活者の口コミによって流行したと「口コミ指数」によって分類された商品・サービスと話題の内容を確認する．実際のブログの記事を読むと,「TSUM TSUM」は，ゲームで高得点を取る方法がブログ上で書きこまれており，そうした内容を求めて人々が検索を行なった結果と考えられる．また，表 4.2 でも示した通り,「アナと雪の女王」は主題歌が評判になり，様々な著名人，一般の個人が歌う動画が投稿され大きな流行となっていた．このように生活者の口コミによって流行した商品・サービスの話題には，生活者が体験した感想や共感できる内容が含まれていることがわかる．一方で，マスメディア上の露出によって流行した商品は，表 4.3 で一部見られたようにマスメディアで接触した話題に対する期待感や，ニュースを拡散しようとするブログ記事が多かった．

表 4.5: 生活者の口コミによって流行になったと思われる商品と話題の組（全 14 組）の一部.

順位	口コミ指数	人気度	期間	商品	主な話題
1	8.32	393.33	2015/03/22〜2015/07/04	TSUM TSUM	めざましテレビで紹介される.
2	8.21	99.57	2014/02/09〜2014/05/24	TSUM TSUM	攻略サイトでの高得点取得方法の紹介.
3	7.60	14.10	2013/05/12〜2013/08/24	あべのハルカス	6/13 に近鉄百貨店開業と報道.
4	6.47	365.30	2014/06/15〜2014/09/27	アナと雪の女王	DVD 発売開始.
5	4.19	635.05	2014/04/06〜2014/07/19	アナと雪の女王	レリゴー現象（主題歌が話題になる）.
6	3.76	203.17	2014/05/11〜2014/08/23	TSUM TSUM	テレビ番組で紹介.
7	3.31	354.32	2014/11/30〜2015/03/14	TSUM TSUM	年末年始でアプリのダウンロードが伸びたため.
8	3.31	85.15	2014/11/30〜2015/03/14	アナと雪の女王	年度末のヒット商品ランキング 1 位の効果.
9	2.76	2.88	2012/07/29〜2012/11/10	あべのハルカス	地上高 300m に到達の発表.
10	2.26	553.15	2014/02/16〜2014/05/31	アナと雪の女王	3/14 公開.
平均	4.13	271.85			

表 4.6: マスメディア上での露出によって流行になったと思われる商品と話題の組（全 18 組）の一部.

順位	口コミ指数	人気度	期間	商品	主な話題
1	0.035	1.86	2013/10/13〜2014/01/25	クロワッサンドーナツ	テレビ番組で流行している商品として紹介される.
2	0.070	1.45	2014/10/05〜2015/01/17	クロワッサンドーナツ	ヒット商品ランキング発表.
3	0.080	4.34	2014/04/13〜2014/07/26	ジェルボール洗剤	誤飲に関する報道.
4	0.102	3.52	2014/06/22〜2014/10/04	ジェルボール洗剤	ヒット商品ランキング発表.
5	0.117	0.747	2014/08/31〜2014/12/13	伊右衛門 特茶	Twitter，FB 1 周年キャンペーン.
6	0.138	15.59	2014/09/28〜2015/01/10	格安スマホ	楽天が参入
7	0.154	0.876	2014/05/04〜2014/08/16	伊右衛門 特茶	Twitter フォローキャンペーンの発表.
8	0.160	40.99	2014/02/02〜2014/05/17	あべのハルカス	3/7 に全面開業と発表.
9	0.184	34.63	2014/03/30〜2014/07/12	あべのハルカス	GW のレジャースポットとして話題になる.
10	0.234	0.284	2013/05/26〜2013/09/07	クロワッサンドーナツ	パン屋で販売されたものを食べた感想.
平均	0.236	11.49			

4.6.2 SIR モデルを利用した分析結果

SIR モデルを利用した全体の分析結果を表 4.7 に示す．なお，表 4.7〜表 4.9 内で $\overline{C_*}$ は，式 4.9 で示したように各メディアでの露出量を予測の時点の前までで平均したものを示している．同一の話題の期間について，SIR モデル近似によるパラメータ推定と各指標の相関分析を行った結果，期間中の人々の『人気度の総和 (S_0)』はブログ・テレビの露出量に正に相関していた．一方で，情報の『拡散力 (β)』,『減衰力 (γ)』はブログ，新聞，テレビの露出量に弱く正に相関する結果となった．この結果は，例えばテレビの場合，露出が増えるほど「テレビでの露出に接することで興味を持つ可能性を持つ人の総数」を増やし，実際に興味を持ちやすくなりつつも，飽きやすくもなる，と解釈することができる．

生活者の口コミが起点となって拡散した情報に関するデータの分析結果を表 4.8 に示す．期間中の人々の『人気度の総和 (S_0)』はブログ，テレビの露出量に正に相関する結果となった．マスメディア上での露出によって拡散した情報に関するデータの分析結果を表 4.9 に示す．期間中の人々の『人気度の総和 (S_0)』はブログ，新聞，テレビの露出量に正に相関する結果となった．情報の『拡散力』,『減衰力』は新聞，テレビの露出量に正に相関する結果となった．

表 4.7: 推定したパラメータと指標との相関係数.（表中の “*” は無相関検定の検定結果で，それぞれ $*: p < 0.05, **: p < 0.01, ***: p < 0.001$ で有意であることを示す.）

パラメータ	指標							
	$\overline{C_{blog}}$	$\overline{C_{twitter}}$	$\overline{C_{nikkei}}$	$\overline{C_{MetaTV}}$	$\overline{log\ C_{blog}}$	$\overline{log\ C_{twitter}}$	$\overline{log\ C_{nikkei}}$	$\overline{log\ C_{MetaTV}}$
S_0	0.955^{***}	0.391^{**}	0.340^{*}	0.761^{***}	0.728^{***}	0.612^{***}	0.344^{*}	0.662^{***}
β	0.378^{**}	0.0946	0.401^{**}	0.349^{*}	0.480^{***}	0.378^{**}	0.329^{*}	0.476^{***}
γ	0.379^{**}	0.0942	0.400^{**}	0.349^{*}	0.481^{***}	0.379^{**}	0.328^{*}	0.477^{***}

4.7 考察

本章では考察を述べる．

4.7.1 商品・サービスの類型化について

ヒット現象は口コミ指数を用いることで，生活者の口コミによって起きるものとマスメディア上の露出によって起きるものの 2 つに類型化することができる．そし

表 4.8: 生活者の口コミによって流行したデータでの推定したパラメータと指標との相関係数.（表中の “*” は無相関検定の検定結果で，それぞれ $*: p < 0.05, **: p < 0.01, ***: p < 0.001$ で有意であることを示す.）

パラメータ	指標							
	$\overline{C_{blog}}$	$\overline{C_{twitter}}$	$\overline{C_{nikkei}}$	$\overline{C_{MetaTV}}$	$\overline{log\ C_{blog}}$	$\overline{log\ C_{twitter}}$	$\overline{log\ C_{nikkei}}$	$\overline{log\ C_{MetaTV}}$
S_0	0.919^{***}	0.109	0.386	0.809^{***}	0.794^{***}	0.512	0.420	0.818^{***}
β	0.522	0.0184	0.351	0.303	0.604^{*}	0.465	0.260	0.449
γ	0.522	0.0192	0.351	0.305	0.607^{*}	0.468	0.260	0.451

表 4.9: マスメディア上での露出によって流行したデータでの推定したパラメータと指標との相関係数.（表中の “*” は無相関検定の検定結果で，それぞれ $*: p < 0.05, **: p < 0.01, ***: p < 0.001$ で有意であることを示す.）

パラメータ	指標							
	$\overline{C_{blog}}$	$\overline{C_{twitter}}$	$\overline{C_{nikkei}}$	$\overline{C_{MetaTV}}$	$\overline{log\ C_{blog}}$	$\overline{log\ C_{twitter}}$	$\overline{log\ C_{nikkei}}$	$\overline{log\ C_{MetaTV}}$
S_0	0.635^{**}	0.424	0.664^{**}	0.697^{**}	0.555^{*}	0.461	0.587^{*}	0.628^{**}
β	0.502^{*}	0.231	0.455	0.479^{*}	0.429	0.314	0.395	0.431
γ	0.502^{*}	0.230	0.453	0.479^{*}	0.429	0.312	0.393	0.431

て，この 2 つの類型について，情報拡散の過程に差異を見出すことができる．例えば，マスメディア上の露出によって起きたヒット現象の例として，表 4.6 内の「伊右衛門 特茶」では，ソーシャルメディアを利用したキャンペーンを利用しているが，マスメディア上の露出に対し，実際のソーシャルメディア上の露出はあまり多くないことが口コミ指数によって観察される．このタイプのヒット現象は，表 4.6 に示した通り人気度の平均が小さい．つまり，マスメディア上への露出のみで大きなヒット現象を作り出すのは難しいことを示している．また，一般的にメディア露出やソーシャルメディア上の露出の後押しをするために，費用をかけて広告出稿を行う方法も考えられる．本章で行う研究ではこうした手法について分析対象とはしていないが，この点についても踏み込んだ分析を行うことで，また異なる知見を得られる可能性もある．

一方で,「アナと雪の女王」のように生活者の口コミによって大ヒットする商品が認められた．このタイプのヒット現象は，表 4.5 に示した通り人気度の平均が大きく大ヒット現象となっているものが多い．生活者の口コミによって流行した事象に関する表 4.8 の β, γ と各メディアでの露出量との相関関係を見ると，特にソーシャルメディアにおいて，$\overline{C_{blog}}$ よりも $\overline{log\ C_{blog}}$，$\overline{C_{twitter}}$ よりも $\overline{log\ C_{twitter}}$ の方が相関が大きいことが確認できる．これは生活者の口コミによる拡散力や減衰力は，メディアでの露出量に対し，対数線形的に相関していることを示唆している．同様の比較をマスメディア上の露出によって流行した事象（表 4.9 の β, γ と各メディアでの露出量との相関関係の比較）でも確認すると，こちらはこのような対数線形的な相関性は確認できない．マスメディア上の露出によって流行する商品は人気度が比較的小さいものが多いことから類推すると，このような相関の線形性の違いは，商品が獲得する人気度の違いから生じるものだと考えられる．つまり，人気度が高い商品を大量にメディアに露出させることで，最初は線形的にその商品に興味を示す人が現れるが，やがては対数線形的になり，ゆるやかに増加するようになる．つまりソーシャルメディア上の露出の数をコストをかけて増やしても効果は相対的に小さくなることが考えられる．

ここで取り上げた事象はすべてヒット商品に関する流行現象であるので，全てヒット現象として取り扱っている．しかし,「アナと雪の女王」もメディア上の露出数は非常に大きく，生活者の口コミのみで流行したわけではない．口コミによる評判の良さとメディアの露出の相乗効果によって大きなヒット現象になっていることについては注意しておきたい．また，メディア間での密接な相互作用については，本章では取り扱っておらず，今後の課題としたい．

また,「TSUM TSUM」に関しては，どの流行現象についても口コミ指数が高いものとなっている．これは，スマートフォン向けのゲームアプリケーションという性質上，スマートフォン上のアプリケーションプラットフォームで，人気アプリと

して多くのユーザに提示されているという原因も考えられる．ユーザが接触するメディアの情報としてスマートフォンの画面の重要性は非常に高まっており，今後の分析が必要と考えられる．

そして，ソーシャルメディアによる「口コミ効果」の重要性は，すでに多くの関係者に認識されている．本章でも，ソーシャルメディアを利用したキャンペーンなどの事例が見られたが，中には関係者がいかにも生活者としてブログ投稿を行う例も話題になっている．本章で行う研究では，ブログの個々の記事について詳細な分類は行っておらず，こうした作為的な口コミ効果を含んでいる可能性があることにも言及しておく．

4.7.2 情報発信手法に関する考察

表 4.7 に示した結果から，ヒット現象の生活者による検索数の総量には，ブログやテレビでの露出数が関係していることがわかった．また，ブログ，新聞，テレビでの露出は，多くの人々に対してより急速に情報を拡散させる上で有効である可能性が示された．

流行した話題の内容ごとのタイプ分けの結果を比較すると，生活者の口コミが元となり流行したタイプについては，新聞，テレビでの露出数が，情報の『拡散力』と密接に相関していることが明らかになった．これは，ソーシャルメディア上で話題となる商品は，魅力的であるので，他の生活者の関心をひきやすく，マスメディアで露出させることによって，より多くの人々へ情報拡散を促進すると考えることができる．つまり，もともと流行するだけのポテンシャルがあり，すでに特定のコミュニティ内に口コミで流行しているものをマスメディアで多くの人々に認知させることで，さらに情報拡散を促すことができることを示している．また,「アナと雪の女王」の事例で見られるように，ソーシャルメディア上で話題の動画をマスメディアに取り上げてもらうことでも，同様の現象が発生する可能性があると考えられる．

一方で，マスメディア上での露出によって流行したタイプではブログと TV の露出量が『拡散力』と一定の相関があることが認められた．ブログや TV での露出は，生活者に一定の関心を持たせるような影響を与える．

ここまでの考察をまとめると，メディアがヒットに与える影響としては，どのメディアでも露出量が多ければ多いほど，人気の合計（検索数の総量）は上昇する．また，ツイッター以外のメディアでの露出は,『拡散力』,『減衰力』に正の相関がある．つまり，これらのメディアでの露出は急速な拡散が見込めると同時に早期の減衰も引き起こす可能性が高まることを示唆している．また，生活者の口コミ発で情報が拡散した商品・サービスは，テレビや新聞で露出させることで，大きな『拡散力』を得ている．一方，マスメディアの露出によって情報が拡散した商品・サービスは，

ツイッター以外のメディアによって，大きな『人気度の総量』を得ることができる．

また，本章で行う研究の成果の制限・限界について言及する．まず，本章では『商品に関心を持ち検索する可能性のある人々の総数 (S_0)』を人気度の総量として扱っている．SIR モデルのパラメータは他にも『拡散力 (β)』,『減衰力 (γ)』というパラメータがある．仮に S_0 が大きい場合でも，β が小さく，γ が大きい場合には，商品に関心を持ち検索する可能性のある人々の総数のうちごく一部のみが購買するのみで，流行が終わってしまう可能性がある．このような現象は，メディア上での露出が限定的で一部の人々のみが接触できた場合に起きると考えられる．しかし今回対象にした商品はすべてヒットした商品であるので，このようなマーケティング不足で流行せずに終わった事象は少ないと考える．また，今回はある商品・サービスについて，同一内容の情報が話題となった期間のみを抽出して分析を行った．しかし，実際のヒット商品・サービスは長期間にわたってヒットが継続しているが，その期間の中で発信する内容を変化させることで，継続して生活者の関心を増大させている．本章での分析では，次々と発信内容を変えることが人気度の変化にどのような影響を与えるかは分析していない．また，同一の話題で情報発信を行われている期間の同定については，各メディアへの情報の露出状況を確認することで行っている．同一の話題の時系列の分布を取得する手法として Dynamic Topic Models[8] といった手法が提案されている．こうした手法を利用し正確に商品の情報拡散の話題を捉えることで，より精緻な分析やモデル構築を行うことができるだろう．

4.7.3 本章の貢献

ここまでの議論から，情報拡散（ヒット現象）の発生にはブログによる情報発信と，テレビを通じた情報発信が比較的大きな影響を及ぼしている可能性が示された．また，ヒット商品・サービスに関する事例を「口コミ指数 (I_{WoM})」を用いて類型化したところ，マスメディアによって流行した商品は全体的に人気度の総量が小さく，キャンペーンや単発のイベント，ニュースに関連するものが多くみられた．一方で，生活者の口コミによってヒット現象となった商品は，人気度の総量が大きく，大きなニュースや生活者が参加しやすい内容のものが多くみられた．また，それぞれのタイプにおいてメディアが及ぼす影響力も異なることから，目指すべきヒットのタイプを明確にし，情報発信手法を設計することが必要なことが明らかになった．

本章で行う研究の成果を要約すると，以下の貢献が挙げられる．

- 人々による商品・サービスへの検索数 (関心) の推移が，ソーシャルメディアやマスメディアの露出量と相関していることを示した．
- 商品・サービスに関する情報が拡散し生活者の関心が増減する過程には,「口コミ指数 (I_{WoM})」を用いることで，マスメディア上の露出によるものと生活

者の口コミによるものに分類でき，それぞれ異なる情報拡散過程を示すことがわかった．

- マスメディア上の露出による情報拡散は人気度が小さく，大量の人気度の獲得には生活者の口コミが重要であることがわかった．

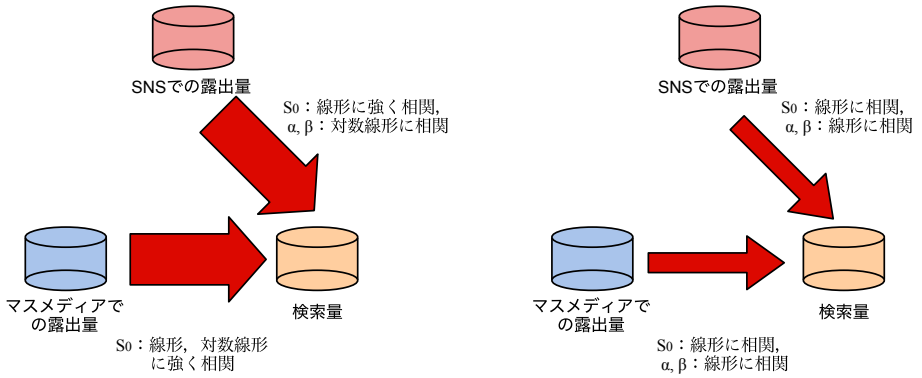
今後，本章で行った研究の成果を利用することで，ある商品・サービスの広報活動の成果を最大化するために，どのようなメディアを活用するべきか指針を得ることが可能になる．一方，今回はある商品・サービスの中でも同一の内容の情報が拡散している期間を同定してデータを抽出して分析を行った．今回の分析をもとに今後は，長期間にわたってヒット現象が続く商品・サービスについても説明可能なモデルを構築し，生活者への影響を拡大する情報発信手法の実証を目指していきたい．

4.8 本章での結論

本章で行った研究ではヒット商品・サービスに関する情報の拡散過程を分析し，マスメディアやソーシャルメディアでの情報の露出によって，人々による該当商品・サービスへの検索数が変化する過程を明らかにした．また，商品・サービスに関する情報拡散（ヒット現象）には，企業発のニュースが起点となるタイプと，生活者の口コミが起点となるタイプとがあり，それぞれのタイプに影響を及ぼしているメディアも異なることが分かった．本章で行った研究の成果によって，インターネット上の情報拡散の重要性が増している現在，商品・サービスの情報を伝えるために広報実務家がどのような情報発信を行うべきか，検討するための一つの指標を提示している．

なお，一つの商品・サービスに関する長期間のヒット現象については，実際には複数の話題による情報発信の結果が組み合わさり，生じている．今回は，複数の話題の相乗作用によるヒット現象の分析は行なっていない．また，単一のメディアの影響のみならず，複数のメディアの影響を同時に折込んだ動的な情報拡散モデルの構築も必要となるだろう．

ここで，本研究全体における本章での結論について述べる．本章では，第 3 章で提案した，異種データ間のユーザ移動ネットワークの複数メディアでの露出量と検索量という実データへの適用と知識抽出の効果の検証を行った．特に本章では，ユーザ移動ネットワークのうち，データソース間でユーザを一意には特定できず，部分的に特定可能な場合の検証を行った．ここでは，ある商品に関する情報だけを選択することによって，その商品に関心を持つユーザ群として特定することで，ソーシャルメディア，マスメディアのデータと検索量のデータ間に因果関係を仮定し，その度合いの検証を行った．検証の結果，ソーシャルメディア，マスメディアの露出量



(a) 口コミ指数が大きいヒット現象　　(b) 口コミ指数が小さいヒット現象

図 4.6: 口コミ指数による情報拡散の違い．

と検索量の推移に相関があることがわかり，大ヒットにはマスメディアの露出量だけでなくソーシャルメディアでの露出量が大きな役割を果たしていることがわかった．ここで，図 4.6 に本章で提案した口コミ指数を用いた，メディア間の情報拡散過程の違いをユーザ移動ネットワークに基づいて示す．本章では，ユーザ移動ネットワークを AISAS モデルを仮定した上で構成している．この点に留意した上でユーザ移動の議論を行う．

まず，口コミ指数が大きいヒット現象の場合について，表 4.7，表 4.8 を合わせ，考察を行うと図 4.6(a) のようになる．この図は以下の内容を示している．ある話題がＳＮＳ上で露出したときに，それを見たユーザが検索を行う数は，ソーシャルメディアでの露出が増えるほど線形的に増加する．そして，その行動は，露出量を増やすほど，対数線形的に比例する形で，より迅速に検索行動を行うようになる．一方で，マスメディア上で露出したときでは，それを見たユーザが検索を行う数は，マスメディアでの露出が増えるほど増加する．しかし，マスメディア上での露出を増やしたところで，ユーザが関心を持ち，検索するまでの期間を短縮することはできないことを示している．

次に，口コミ指数が小さいヒット現象の場合について，表 4.7，表 4.9 を合わせ，考察を行うと図 4.6(b) のようになる．この図は以下の内容を示している．ある話題がＳＮＳ上で露出したときに，それを見たユーザが検索を行う数は，ソーシャルメディアでの露出が増えるほど線形的に増加する．そして，その行動は，露出量を増やすほど，線形的に比例しより迅速に検索行動を行うようになる．一方で，マスメディア上で露出したときでは，それを見たユーザが検索を行う数は，マスメディアでの露出が増えるほど増加する．また，マスメディア上での露出を増やすと，ユーザが関心を持ち，検索するまでの期間を短縮することもできる．

以上のようにユーザ移動ネットワークによって，メディア間の情報拡散過程の違

いを示すことができた．これによって，第 3 章で提案した，ユーザ移動ネットワークのうち，データソース間でユーザを一意には特定できず，部分的に特定可能な場合について，実際の複数のデータに適用し，有用な知識を抽出できることを示した．本章での検証によって，ユーザ移動ネットワークを適用することによって，獲得する知識の価値が高まることが示唆されている．

第5章　スマートカード上の大規模移動データを利用した移動目的推定モデルに基づいた地理的地域の分散表現の獲得

5.1　本章で用いるユーザ移動ネットワーク

第3章では，異種データ間のユーザ移動ネットワークに基づいた複数公共データの統合による知識抽出手法を提案し，定式化を行った．そこでは，ユーザ移動ネットワークの作成にはデータソース間でユーザを一意に特定できる場合と部分的に特定できる場合があり，一意に特定できる場合には，データソース間の関係性に仮定を置かず直接ユーザ移動ネットワークを形成できることを述べた．本章では，データソース間でユーザを一意に特定できる場合のユーザ移動ネットワークの具体例としての分析を行い，知識を獲得する．本章で用いる異種データ間ユーザ移動ネットワークの詳細な図を図5.1に示す．ここでは，公共交通における乗降履歴に関するデータを用いる．この乗降データは，複数の運営会社がそれぞれに管理するデータである．これらのデータはスマートカードとよばれる非接触型ＩＣカードを通して記録されるもので，それぞれの運営会社が管理するデータソースをまたいでもユーザＩＤは共通となっているため，データソース間で一意にユーザを特定することができる．図中で社局を超えた移動は，多くの場合，ユーザは徒歩で乗り換えを行っている．この移動量は，ユーザを一意に特定できるために算出することができる．本章ではこのデータのうち，各駅ごとの乗降に注目し，ユーザ移動ネットワークを図に示したように適用する．そして，ある駅に行く理由をこのユーザ移動ネットワークから推定することを目的とする．特にユーザを一意に特定できるデータの場合，そのユーザの移動の系列情報（以降，文脈情報と呼ぶ）を利用することが知識抽出に効果的であることを示す．

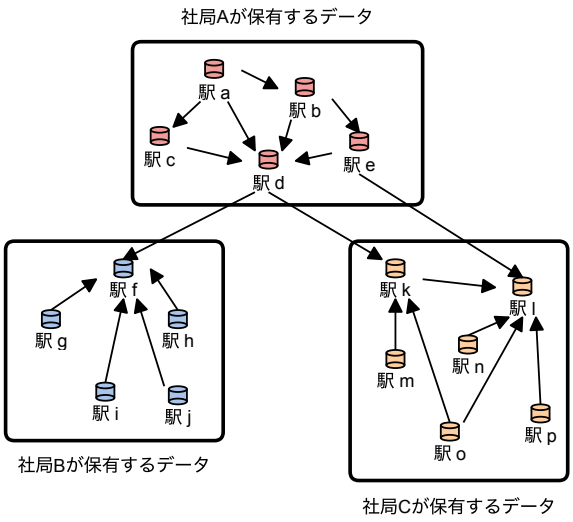


図 5.1: 第 5 章で用いる異種データ間ユーザネットワーク.

5.2 本章の背景，目的

位置情報のセンサー装置とネットワークが広く普及するに連れて，人々の移動データが大量に蓄積されるようになり，しかも研究に利用できるようになってきている [81, 83]．このセンサー装置に加えて，スマートカードを利用する自動料金収受システム (AFC) のような公共交通基盤の発達は，詳細な時間や位置情報を含む大量の人々の移動データの蓄積を可能にした [60].

この大量の移動データを位置情報に基づいた個人化された興味のある場所 (PoI) の推薦の目的で利用した研究がなされている [75]．最近では，地域開発 [43] や都市開発 [82]，政策決定 [32] のために移動データを用いようという研究がある．これらの研究の興味は，移動データが集積した地域で，どのように人々の移動をモデル化し，予測するかということである．

特定の地域での人流のモデル化と予測は，その地域についての外部情報と人々の活動データを結合することによって，その地域の特徴や果たしている役割に対する理解を可能にすると考えれる [53]．最近の研究では，スマートカードのデータを使ってある地域から別の地域への人々の移動パターンを分析し，地域を特徴化したり，地域をいくつかに分類するという取り組みがされている [44, 78]．ただし，これらの研究は単にその地域の人流に基づいて地域の分類を行い，ある地域は事前に定義された属性のいずれかになることを仮定している．しかし，ある地域の大規模な人々の移動パターンをその地域における文脈であるとみなせば，その地域の特徴や役割が，このような静的なものではなく，その地域の人々がどう移動しているか，またその地域に人々がどういう目的で訪問するのかを考えれば，動的に変化していることに気づくだろう．言い換えると，２つの地域が類似した特徴や役割を持っている

ならば，それらはそのような文脈によって定義されうる共通の地域の表現を持つはずである．もし，地域の潜在的な表現を獲得することができるならば，それは巨大な移動データにある人流をもっと効果的に正確にモデリングしたり予測することに貢献することができるだろう．

こうした文脈情報を利用した表現学習の基本的な概念 [7] は，2 つの実体が共通の文脈を共有しているならば意味的に類似しているというものだ．これは，言語学において分布仮説として知られており，そこでは同様の文脈で生起する語群は類似した意味を持つ傾向があることが説明されている [47]．文脈情報を利用した表現学習におけるこの考え方は最近，ネットワーク表現学習法 [11, 52, 64] へと拡張されてきている．ネットワーク表現学習法はネットワーク内で関係性の近い 2 つのノードを類似していると仮定することによって低次元ベクトル空間へネットワークをエンベディングしようとする試みである．地理的な地域の潜在的な表現を獲得しようとする場合，地域をノード，地域間の人々の移動パターンをリンクとみなすことによって，この課題をネットワーク表現学習法の研究の拡張として定式化できる．この問題を直感的な例で説明すると，商業地域の人々の移動パターンと居住地域の人々の移動パターンとでは異なっており，移動パターンを人で見るだけでも簡単に区別できるだろうというものだ．そこで本章では，この人々の移動ネットワークを低次元ベクトル空間へエンベディングし，獲得した分散表現を元に地理的な地域のタイプの違いを区別する．

本章での研究では，この文脈情報を用いる表現学習の技術を利用して地理的な地域の潜在的な表現を発見することを目指す．獲得した表現は，地理的な地域や関係性の持つ潜在的な役割を明らかにすることができるので，都市計画や地域開発に利用できる可能性がある．こうした役割は，移動データの表面的な情報のみからでは，簡単には観察できないようなものである．本章では，この分散表現を大規模な人流データから獲得するために既存のネットワーク表現学習の概念（ネットワークエンベディング法）を適用する．しかし，既存のネットワークエンベディング法を本章での地理的な地域をエンベディングしようとする問題へ単純に適用することはできない．例えば，鉄道のような巨大交通システムネットワークのおける人々の移動には，地理的な制約というものが存在する．これは，東京に住んでいる人は日用品を買いにわざわざ大阪へは行かず，日用品は家の近くでいつも買い，ちょっとしたことではるばる遠くへ行ったりはしないという行為に端的にあらわれている．この例でもわかる通り，人々は現在の居場所と利用できる交通手段に応じて移動距離を最小化しようとする傾向があることが仮定できるのではないだろうか．そこで，本章ではこうした地理的な制約を盛り込み，“**移動目的推定モデル**” として提案する．これは，地理的な地域を地域間を移動する人々の移動パターンで接続するネットワークとして考え，そしてそのネットワークを低次元ベクトル空間へ写像しようとした場合に，

実世界上の人々が移動する上での地理的な制約を考慮する必要がある，というものである．この移動目的推定モデルを表した図を図 5.2 に示す．この図は，人流グラフ (G_{ss}:図 5.2 右)，地理的制約グラフ (G_{sc}:図 5.2 左)，目的近接性グラフ (G_{sr}:図 5.2 中) の 3 つから構成されている．人流グラフは人々の駅間の移動量を表すグラフである．地理的制約グラフは，駅間の地理的な近接性を表したグラフであり，図 5.2 では路線図をあげている．目的近接性グラフは，ある駅へ移動する目的と駅とを頂点とし，ある駅とその駅へ移動する目的とをエッジとして接続したグラフである．通常このある地域に移動する目的は移動したユーザ自身に聞かなければわからないだろう．その結果を集めるにはコストがかかるため，十分な量のデータを取得できていることは少ないと考えられる．そこで本章で提案する手法では代表的な少数の駅での目的近接性グラフのみを利用し，その他の駅に行く目的を推定する，というタスクを設定する．モデルでは，地理的制約と目的の 2 つから人々は移動する目的地を決定するという仮説にしたがって，半教師あり学習の枠組みに基づいて，駅へ移動する目的を推定する．つまり，本章では，大規模に蓄積されたスマートカードの利用履歴に基づいた人流データから抽出された人々の移動パターンを使って，地理的な地域を訪問する目的の分散表現を獲得するためのエンベディング法を提案する．提案手法は 2 つのエンベディングモデルから構成される．それは “**接続モデル (concatenating model)**” と “**内分モデル (internally dividing model)**” である．本章では，関西地方の大規模な鉄道路線ネットワークの大規模なスマートカードの利用履歴を使って実験を行った．提案した移動目的推定手法を使用してそれぞれの鉄道駅の分散表現を獲得し，それを鉄道駅に対する複数ラベル分類のタスクを行うことで評価を行った．本章ではこの実際の大規模移動データ上で提案手法がうまく機能することを示す．提案手法は，地理的には離れていても同様の特徴や役割を持っている鉄道の大規模ネットワーク上の駅を識別することができる．そして，地理的な地域や関係性の持つ潜在的な特徴を示すことで都市計画，マーケティング，政策立案の担当者が戦略を設計したり，政策を地域開発のために実行するのを助けることができると考えれる．

本章の貢献は以下の 4 つに要約できる．

1. 移動目的推定モデルを提案し，人々の移動パターンを使用する地理的な地域の分散表現を獲得する新しいエンベディングモデルを開発した．

2. 提案したモデルが国内の鉄道のスマートカードによる実際の大規模移動データを使ってうまく機能することを示した．

3. 提案したモデルが地理的には離れていても同様の特徴や役割を果たしている駅を識別できることを示した．

4. 提案したモデルのパラメータ推定の結果によって，移動目的の方が地理的な近

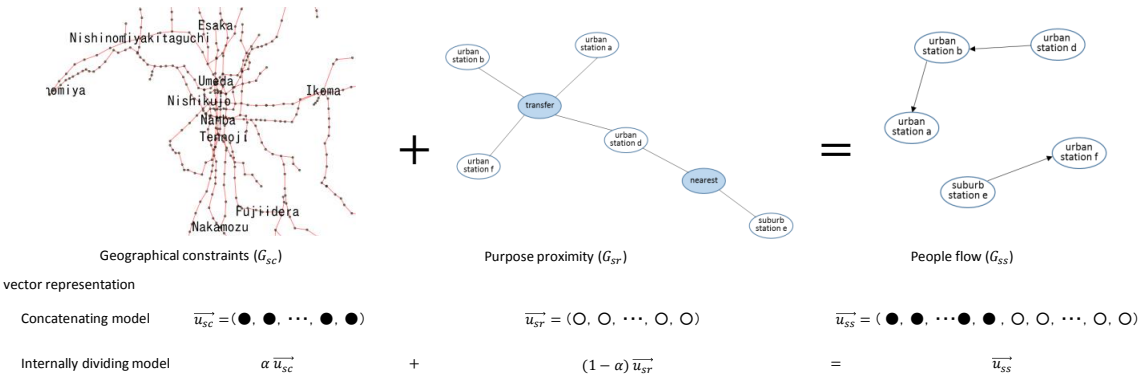


図 5.2: “移動目的推定モデル” の概念図.

接性よりも 1.1 倍重要であることがわかった.

5.3 関連研究

本章で行う研究は移動データ分析とネットワーク表現学習に関する研究と関係が深い．本節では，本章で行う研究の位置づけと既存の研究との関係を議論し，新規性と意義について明らかにする.

5.3.1 移動データを用いた地理的な地域の特徴のモデリング

最近のスマートカードによる自動料金収受システム (AFC) のようなセンサーネットワークと公共交通の基盤は詳細な時刻と場所の情報を持つ人々の活動を含んだ大量の移動データの集積を可能にしている．特に，AFC システムからの移動データは可視化 [25]，障害予防 [74]，サービス管理 [36] に利用されている.

さらに，地点に対する興味 (PoI) の個人への推薦 [75]，地域開発への推薦 [43]，都市計画への推薦 [82]，政策策定への推薦 [32] に対する応用のために，さまざまな研究が特定の地域における人流をどのようにモデル化するか，大規模移動データで地域の特徴をどう理解するかということに取り組んでいる.

ある地域から別の地域への人々の移動パターンを分析し，地域を特徴づけたり，地域の類型化を行ったりする研究がなされている [44, 77]．これらの研究は，単にある地域がその地域の人流に基づいて複数の事前に定義された属性へ分類されることを仮定している．しかし，ある地域での人々の大規模な移動パターンをその地域の文脈と捉えた場合，地域の特徴や役割がその地域でどのように人々が移動しているか，人々が何のためにその地域を訪れているのか，と言った文脈によって刻々と変化していることに気づく．本章では，エンベディング手法を用いてそのような文脈

によって定義される地域の共通の本質的な表現を獲得することを目指す．地域の特徴や役割を理解する場合，過去の研究は商業地区や住居地区のような事前定義された属性情報を必要としている [56, 82]．一方で，本章で提案するモデルは半教師有り学習の枠組みで，こうした情報を少数のタグ付けされた地域情報から学習することができる．

5.3.2 ネットワークデータのエンベディング

本章で行う研究において，文脈情報を用いた表現学習の技術を利用して地理的な地域の潜在的な表現を発見することを目指す．本章で提案する手法は，既存のネットワークエンベディング法をベースにした手法を大規模人流データからこのような表現を獲得するために適用する．このネットワークエンベディング法はグラフ理論と言語学的な単語エンベディング法に由来している．グラフ理論の文脈では，特異値分解 (SVD) や非負行列分解 (NMF) のような隣接行列の行列分解技術がプロトタイプである [15]．一方で，単語エンベディング法が最近発展してきている．単語エンベディング法の基本的な概念は，２つの実体が共通の文脈を持っている場合，それらは意味的に類似しているというものだ．これは言語学では分布仮説として知られており，同様の文脈で生ずる単語群は類似した意味を持つ傾向があるというものである [47, 40]．グラフ構造を直接エンベディングしようとした研究も行われた [11, 52, 64]．これらのネットワークエンベディング手法は可視化，ノードの分類，リンク予測のような多くのタスクで有用であることが示されている [80]．

ネットワークエンベディングは時系列解析 [34] や異種ネットワーク [63] に対しても開発されている．特に，PTE 法は異種ネットワークのエンベディング法で，単語，文書，ラベルに関するネットワークを元に低次元のベクトル空間へ写像する．PTE 法はこれら３つの異種ネットワークを同一のベクトル空間に写像する．本章で提案するモデルも同様に３つの異種ネットワークをベクトル空間に写像するが，それらを別々のベクトル空間へ写像する点で異なる．以下の章からは，これについて詳細に述べる．

5.4 提案手法

本節ではまず人流は地理的な場所と目的から生じるという “**移動目的推定モデル**” について説明する．そして，大規模な人流からどのようにネットワークが形成されるのかということと，ネットワーク上のラベル伝搬の必要性について説明する．さらに仮説に基づいたモデルを提案し，詳細に説明を行う．本章で提案する手法は，ラ

ベル伝搬ネットワークエンベディングモデルを地理的な制約のある大規模人流データのために拡張したものである．

5.4.1 移動目的推定モデル

本項では，GPS データ，携帯基地局のデータ，鉄道の乗降データ，等のような人流データのための，図 5.2 に示すような “移動目的推定モデル” を提案する．このモデルは，ネットワーク上の各頂点をベクトル空間に写像する手法の一つである．提案するモデルは，３つのネットワークが存在している状態を想定している．それは，人流ネットワーク（図中 : 右）（人々の移動量を表すネットワーク），目的近接性ネットワーク（図中 : 中）（地域とそこに目的とのつながりを表すネットワーク），地理的制約ネットワーク（図中 : 左）（路線図など，地域間の近接性を表すネットワーク）である．これら３つのネットワークは同一の地域が頂点となっている．人流ネットワークは出発地から目的地への移動の割合（$P(destination\ area|departure\ area)$）で辺の重みを設定した有向グラフである．地理的制約ネットワークは，地域の近接関係を同一の重みで設定した無向グラフである．目的近接性ネットワークはある地域に行く目的の割合（$P(purpose|destination\ area)$）で辺の重みを設定した無向グラフである．特に，地理的制約ネットワークと人流ネットワークは地域のみで構成されている．一方で目的近接性ネットワークは，地域と目的を頂点とし，地域とその地域へ行く目的を辺として接続したネットワークである．本項で提案するモデルはこの地域，目的を表す頂点をそれぞれのネットワークを表す別々のベクトル空間へ写像を行う．つまり，同じ地域でも３つのネットワークで異なるベクトル空間へ写像されている点で，PTE 法と異なる．そして，この目的近接性ネットワークは地域へ行く目的が代表的な地域に限り少量判明している場合を想定している．本項で提案するモデルは，半教師有り学習の枠組みを用い，地理的制約ネットワーク，人流ネットワーク，目的近接性ネットワークの３つを利用して，目的近接性ネットワークに存在しない地域もベクトル空間に写像を行う．

本章では，スマートカードシステムから抽出された関西地域の鉄道の乗降データを人流データとして適用する．地域は図中の駅として表現することとする．このモデルは，人は今いる場所では達成できない目的を達成するためにどこかへ移動するというモデルである．言い換えると，人々の移動 (人流) は地域 (地理的制約) と地域の持つ役割 (目的の近接性) の合計で表されるというものである．そして，モデル上の人々は現在の所在地から目的を達成できる近くの場所へ移動する．この人々の移動とその目的がデータとして蓄積していくにつれて，人流ネットワークが形成される．図 5.2 は，２つのネットワーク（地理的制約ネットワーク（左）と目的ネットワーク（中））から大規模人流ネットワーク（右）が生成されていることを示してい

る．そしてこのような 3 つのネットワークの関係は潜在的な分散表現上の距離に依存していると考える．

相互に分散表現を共有しない 3 つのグラフがあるとしよう．それは，人流グラフ (G_{ss}:図 5.2 右)，地理的制約グラフ (G_{sc}:図 5.2 左)，目的近接性グラフ (G_{sr}:図 5.2 中) である．より具体的には，それぞれのグラフ内の分散表現は以下の通りである．$v_i \in G_{ss}$ の各頂点に対する分散表現は $\vec{u_i^{ss}}$，$v_i \in G_{sc}$ の各頂点に対する分散表現は $\vec{u_i^{sc}}$，$v_i \in G_{ss}$ の各頂点に対する分散表現は $\vec{u_i^{ss}}$ とする．図 5.2 で示したモデルの仮説はベクトル間に成り立つ以下の等式を導く．

$$\vec{u_i^{ss}} = \vec{u_i^{sc}} + \vec{u_i^{sr}} \tag{5.1}$$

この等式を 2 つの形式で解釈する，すなわち “**接続モデル (concatenating model)**” と “**内分モデル (internally dividing model)**” である．“接続モデル” では，“+” 演算子を 2 つのベクトル各要素の和を取るのではなくて，2 つのベクトルを結合し，2 倍の次元を持った新しいベクトルを作ると解釈をする (図 5.2 の Concatenating model)．さらに，“内分モデル” では，式 5.1 を人流グラフのノード ($\vec{u_i^{ss}}$) が地理的制約のノード ($\vec{u_i^{sc}}$) と目的近接性グラフのノード ($\vec{u_i^{sr}}$) の間に位置すると解釈する (図 5.2 Internally dividing model)．これらの 2 つのモデルを以下で説明する．これらのモデルにおける更新式の導出は，第 8 章（付録 A）にて詳細な説明をおこなっている．

5.4.2 接続モデル

接続モデルでは，分散表現は表 5.1 に示した学習アルゴリズムによって獲得される．このアルゴリズムは地理的制約ネットワーク (G_{sc})，目的近接性ネットワーク (G_{sr})，人流ネットワーク (G_{ss})，サンプリングの回数 (T)，初期学習率 (ρ_0)，負例サンプリングの回数 (K)，分散表現の次元数が入力として必要である．Tang らによって提案された “LINE(2nd) model” [64] と呼ばれるネットワークエンベディングモデルを適用する．このモデルは 2 つの頂点間の 2 次の近接性をそれぞれの分散表現を最適化することで近似する．目的関数は以下のとおりである．

$$O = -\sum_{(i,j)\in E} w_{ij} \ln p(v_j|v_i) \tag{5.2}$$

この等式において，w_{ij} は観測可能な頂点 v_i から頂点 v_j へのエッジの重みを表している．頂点 v_i から頂点 v_j への遷移確率である，$p(v_j|v_i)$ は頂点 v_i の分散表現の $\vec{u_i}$ と頂点 v_j の文脈表現の $\vec{u'_j}$ を使って以下のように表される．

$$p(v_j|v_i) = \frac{\exp(\vec{u'}_j^T \cdot \vec{u}_i)}{\sum_{k'\in|V|} \exp(\vec{u'}_{k'}^T \cdot \vec{u}_i)} \tag{5.3}$$

表 5.1: 接続モデルにおける学習アルゴリズム.

接続モデルにおける学習アルゴリズム.	
1:	**Input**: $G_{sc}, G_{sr}, G_{ss}, T, \rho_0, K, d$.
2:	**Output**: $\vec{u^{sc}}, \vec{u^{sr}}, \vec{u^{ss}}$.
3:	それぞれのベクトル $\vec{u^{sc}}, \vec{u^{sr}}, \vec{u^{ss}}, \vec{u'^{sc}}, \vec{u'^{sr}}, \vec{u'^{ss}}$ の初期化.
4:	**for** $t = 1$ **to** T
5:	エッジ e_{ij}^{sc} を G_{sc} からサンプリング.
6:	$\vec{u_i^{sc}}$ と $\vec{u_j'^{sc}}$ を一致するノード $\vec{u_i^{ss}}$ と $\vec{u_j'^{ss}}$ の該当部分から読み出す.
7:	目的関数 O_{sc} を利用して $\vec{u_i^{sc}}$ と $\vec{u_j'^{sc}}$ を更新する.
8:	$\vec{u_i^{ss}}$ と $\vec{u_j'^{ss}}$ の該当部分を $\vec{u_i^{sc}}$ と $\vec{u_j'^{sc}}$ で上書きする.
9:	エッジ e_{ij}^{sr} を G_{sr} からサンプリング.
10:	$\vec{u_i^{sr}}, \vec{u_j'^{sr}}$ を一致するノード $\vec{u_i^{ss}}$ と $\vec{u_j'^{ss}}$ の該当部分から読み出す.
11:	目的関数 O_{sr} を利用して $\vec{u_i^{sr}}$ と $\vec{u_j'^{sr}}$ を更新する.
12:	$\vec{u_i^{ss}}$ と $\vec{u_j'^{ss}}$ の該当部分を $\vec{u_i^{sr}}$ と $\vec{u_j'^{sr}}$ で上書きする.
13:	エッジ e_{ij}^{ss} を G_{ss} からサンプリング.
14:	目的関数 O_{ss} を利用して $\vec{u_i^{ss}}$ と $\vec{u_j'^{ss}}$ を更新する.
15:	END

この目的関数を 3 つのネットワークに個別に設定し，それらを微分することによってそれぞれの頂点ベクトル $(\vec{u_i})$ とそれぞれの文脈ベクトル $(\vec{u_i'})$ の更新式を導出する．最初に乱数によって初期化を行った頂点ベクトルを順番に (7,11,14 行目のように) 接続モデル (8,12 行目のように) に基づいて SGD(確率的勾配降下法) による学習アルゴリズムを使って獲得する．

5.4.3 内分モデル

“内分モデル” では，人流グラフの頂点ベクトル $(\vec{u_i^{ss}})$ は，地理的制約グラフの頂点ベクトル $(\vec{u_i^{sc}})$ と目的近接性グラフの頂点ベクトル $(\vec{u_i^{sr}})$ の間に以下の等式の通りに位置する．

$$\vec{u_i^{ss}} = \alpha \vec{u_i^{sc}} + (1 - \alpha)\vec{u_i^{sr}} \tag{5.4}$$

この等式は，物理的な位置関係と人々がそこで達成したい目的の両方を考慮して目的地を決定することをモデル化している．

目的関数をそれぞれのグラフに対して式 5.2 のように設定する．しかし，グラフ G_{ss} 内の頂点ベクトル $\vec{u_i^{ss}}$ を更新する場合，頂点ベクトル $\vec{u_i^{ss}}$ は頂点ベクトル $\vec{u_i^{sc}}$ と頂点ベクトル $\vec{u_i^{sr}}$ に式 5.4 を通して依存する．そのため，ベクトルとパラメータ α の

ために新しい更新式を導出する必要がある．人流グラフ (G_{ss}) における目的関数 O_{ss} に対し，注意深くすべての依存する変数 ($u^{\vec{l}sc}, u^{\vec{l}sr}, u^{\vec{s}c}, u^{\vec{s}r}$) とパラメータ ($\alpha$) に対して微分を行う．これによって，人流グラフに対し，以下の更新ルールを導出できる．頂点ベクトルと α についての更新式のみを示す．

$$\frac{\partial O_{ij}}{\partial u_i^{\vec{s}c}} = (1-\sigma_{j'i})\alpha u_i^{\vec{s}s} - \sum_{k=1}^{K} \sigma_{k'i}\alpha u_i^{\vec{s}s} \tag{5.5}$$

$$\frac{\partial O_{ij}}{\partial u_i^{\vec{s}r}} = (1-\sigma_{j'i})(1-\alpha) u_i^{\vec{s}s} - \sum_{k=1}^{K} \sigma_{k'i}(1-\alpha) u_i^{\vec{s}s} \tag{5.6}$$

$$\begin{aligned}\frac{\partial O_{ij}}{\partial \alpha} = (1-\sigma_{j'i})\{&(u_j^{\vec{l}sc} - u_j^{\vec{l}sr})^T \cdot u_i^{\vec{s}s} - u_j^{\vec{l}ss^T} \cdot (u_i^{\vec{s}c} - u_i^{\vec{s}r})\} \\ -\sum_{k=1}^{K} \sigma_{k'i}\{&(u_k^{\vec{l}sc} - u_k^{\vec{l}sr})^T \cdot u_i^{\vec{s}s} - u_k^{\vec{l}ss^T} \cdot (u_i^{\vec{s}c} - u_i^{\vec{s}r})\}\end{aligned} \tag{5.7}$$

これらの等式において，$\sigma_{j'i}$ は，$\sigma_{j'i} = 1/(1+\exp(\vec{u_j}^T \cdot \vec{u_i}))$ を表している．また学習アルゴリズムは PTE(joint) 法 [63] と同様のものを適用している．

5.5 入力するデータ概要

本章で説明したように，提案したモデルは 3 つのネットワークデータを必要とする．それは，人流ネットワーク，地理的制約ネットワーク，目的の近接性ネットワークである．これら 3 つのネットワークを入力として整形するために，実験のための 3 つのデータセットを適用する．本節では，これらのデータセットとその整形について説明する．

表 5.2: 乗降データセットの概要.

開始	2013-04-01
終了	2013-04-30
全レコード数	68,763,457
ユニークユーザ数	3,679,251
駅種類数	672
社局数	6

人流ネットワークのための乗降データセット: このデータセットは国内の関西地方

表 5.3: 目的データセットの概要.

目的	のべ人数
通勤	1,278,288
通学	349,234
帰宅	2,192,826
仕事	358,891
その他	1,313,767
合計	5,493,006
駅種類数	599

の大規模なスマートカードデータを含んでいる．このデータセットは関西の鉄道運営事業者 6 社局から提供された乗客のスマートカードのログである．このデータセットは提供事業者によって秘匿化されている．データセットは主に 6 つの要素から構成される．それぞれのユーザの性別，年代，乗車時刻，降車時刻，乗車駅，降車駅である．このデータセットの概要を表 5.2 に示す．本章の研究では人流ネットワークをこの 2 つの駅間を人々が乗降したことを表すデータセットを利用して作成する．有向グラフでそれぞれのエッジの重みは，$P(destination\ station|boarding\ station)$ で表す．今回，朝の出勤の目的を捉えるために 2013 年 04 月の平日の午前 07 時から午前 10 時までの移動データだけを選択した．全日のデータではなく，午前中のデータのみを選択した理由は，乗降データが持つ性質による．多くの人は朝，住居を出発し，どこかに出かけ，帰宅をする，という行動パターンを持っている．つまり，全日のデータを用いた場合，この乗降ネットワークのノードを駅とし，エッジを移動量としたグラフ行列は対称行列に近い形になることが想定される．一方で無向グラフは対称行列として記述することが可能である．本章では，有向グラフを対象としているので，グラフの特徴を明らかにするためにこのようなデータの選択を行った．

地理的制約ネットワークのための路線図データセット: 本章では，路線のネットワーク情報を地理的な近接性を表す情報として適用する．このデータは日本の鉄道路線の API を通して取得した[1]．このデータを通して，路線図を作成した．グラフは無向グラフですべてのエッジの重みは等しい．作成した路線図を図 5.2 の左 (geographical constraints) に示す．

目的近接性ネットワークのための利用目的データセット: 本章はそれぞれの駅へ移動する目的を推定することを目的としている．ここで説明するように，利用目的データセットはパーソントリップ調査の結果を利用してデータセットを作成する．日本

[1] http://www.ekidata.jp/

では，国土交通省が 10 年ごとにたくさんの人々にアンケートを通して国家的に調査を行う．2010 年の結果 [2] を実験へ適用する．そのデータはどのくらい量の人々が個別の駅へどういう目的で行くかというデータを含んでいる．それぞれの駅で降車する目的は，“通勤”，“通学”，“帰宅”，“仕事”，“その他” である．このデータセットの概要を表 5.3 に示す．このデータセットから，駅–目的グラフをある駅へ行く目的の割合の分布であらわす目的近接性ネットワークとして作成する．このグラフは無向グラフである．前述した通り，人流ネットワークを作る時，平日朝のデータのみを選択しているため，このデータセット内の “帰宅” の目的を利用せず，残った 4 つの目的のみを利用する．というのは，ほとんどの人々は朝に帰宅しないと考えられるからだ．

5.6 実験と結果

本節では，開発したモデルの地理的なデータに対する効果を測定する．この目的のため，様々なアルゴリズムと比較する実験を行った．その結果を以下で説明する．

5.6.1 実験手順

本章で説明したように，ある駅で降車する乗客の目的が複数あることを考慮して複数ラベル分類の実験を行った．実際，その駅に降り立つ目的は人によって異なるだろう．駅での降車目的をいくつかの目的の確率分布とみなし，その分布を実験において推定する．

実験の手順は以下のとおりである．まず，5.6.2 で説明するいくつかの手法を用いて分散表現を獲得する．次に，利用目的データセットの一部から作られた訓練用のラベルデータセットを使ってそれぞれの実験用の分類器を訓練する．最後に，テストデータを利用した予測を残りのデータセットから作られたテストデータで行い，いくつかの評価指標を用いて取得した結果の評価を行う．

複数ラベル分類のために，複数クラスのロジスティック回帰分類器を利用した．LIBLINEAR パッケージ [3] を分類器として採用した．分類結果の評価に 3 つの使用を利用する．それは，“KL 距離”，“平均逆順位 (MRR)”， “平均平均適合率 (MAP)” の 3 つである. 実験では，5.5 章で説明した 4 つのクラスの分類をする．各手法の精度を 2 交差検定をデータをランダム化して 5 回繰り返すことで評価を行う．正確に言うと，分散表現を得るためにすべての乗降データを用いるが，一方で，利用目的データセットは半分だけをベクトル表現の獲得と分類器の訓練に用いる．そして残

[2] https://www.kkr.mlit.go.jp/plan/pt/data/pt_h22/index.html
[3] https://www.csie.ntu.edu.tw/∼cjlin/liblinear/

りを分類器の精度の評価に利用する．この実験手順を利用目的データセットをランダム化することで 5 回繰り返して実行する．

最後に，目的ベクトル周辺の地理的な位置の評価を行う．評価指標はある目的ラベルのベクトルの周辺にある駅の実際の地理的な位置の標準偏差の平均値とした．目的ラベルの周辺の駅の標準偏差の平均が大きい場合，その駅群は地理的な制約と無関係に抽出されていることになるからである．

5.6.2 比較アルゴリズム

提案手法の性能の比較に以下のアルゴリズムを利用する．本章での提案手法は，文脈情報を用いた分散表現獲得手法と関係が深いので，それらのさまざまな手法との比較を行った．

1. **重み付きランダム**: 個別の確率分布からランダムにサンプリングする．事前に訓練データからそれぞれの目的の分布を計算する．テストデータである駅の降車の目的を予測する場合，この手法は計算した分布にしたがってランダムに目的を選択する．

2. **Word2Vec** [47]: Word2Vec は大規模コーパスのそれぞれの単語の表現を学習する効率的な単語エンベディングモデルである．本実験では Skip-gram モデルを単純に適用する．

3. **GloVe** [51]: GloVe は別の単語エンベディングモデルである．コーパス全体の単語同士の共起情報を単語表現ベクトルを学習するために利用する．

4. **DeepWalk** [52]: DeepWalk はネットワークの表現を学習することができる最初のネットワークエンベディング法である．このモデルは重みなしのグラフにのみ適用できる．それぞれの頂点にグラフ構造を順序のある配列に変換するために，回数を指定したランダムウォークを実行する．

5. **LINE** [64]: LINE は別のネットワークエンベディング法である．LINE は 1 次の近接性と 2 次の近接性を頂点間のエッジの重み情報を利用して定義する．LINE は頂点ベクトルと文脈ベクトルの内積でそれぞれの近接性に近似することによって表現を獲得する (LINE(1st) and LINE (2nd))．LINE は 1 次の近接性と 2 次の近接性のベクトル表現を接続した場合に最も良い結果を示すことが報告されている (LINE(concat))．

6. **PTE** [63]: PTE は異種ネットワーク向けのネットワークエンベディング法である．この手法は 3 つの異なるネットワーク，すなわち，単語–単語ネットワー

ク，単語–文書ネットワーク，単語–ラベルネットワークに対して適用する．彼らは “pre-train” と “joint” という 2 つの学習形式を提案している．本節では，彼らの報告によると pre-train 学習形式よりもわずかに性能が良い “joint” 学習形式を選択する (PTE(joint))．この手法は 3 つのネットワークグラフ内の頂点を同一のベクトル空間へ写像する．別々のグラフ内の同一のノードは，すべてのグラフ間で同一のベクトル表現で表される．

7. **proposed**: 提案したモデルはすべてスマートカードからの大規模移動データを通して地理的な地域のエンベディングを学習するものである．第 5.4.1 節で説明した “移動目的推定モデル” に基づいた 2 つのモデルを提案する．それは接続モデル (“concat”) と内分モデル (“divide”) である．提案したモデルは 3 つのネットワークグラフの頂点を別々のベクトル空間に写像する．異なるグラフの単一のノードはそれぞれのグラフで異なるベクトル表現 $(\vec{u_{sc}}, \vec{u_{sr}}, \vec{u_{ss}})$ を持つ．

Word2Vec と GloVe は単語エンベディング法であるので，入力として文章が必要である．それぞれのユーザの乗降の履歴である駅の時間的な利用の順序を文章とみなす．Word2Vec，GloVe，DeepWalk，LINE は教師なし学習を行う．そのため，訓練では単に駅の乗降に関するユーザ情報のみを適用する．PTE と提案モデルは半教師あり学習を行う．PTE では，人流ネットワークを単語–単語ネットワーク，地理的制約ネットワークを単語–文書ネットワーク，目的近接性ネットワークを単語–ラベルネットワークとして設定する．すべての手法において，ノードベクトルは 200 に設定し，提案手法の接続モデルでは，ベクトル $\vec{u_i^{ss}}, \vec{u_i'^{ss}}$ は 2 倍の次元，すなわち 400 となっている．

5.6.3 結果

本節では，提案モデルの性能と特徴を明らかにする．

(1) 複数ラベル分類の結果

表 5.4 に複数ラベル分類の結果を示す．まず，重み付きランダム (weighted random) と他の手法との結果との比較から始めよう．重み付きランダム (weighted random) 以外のすべての他の手法は単語やノードをベクトル空間へエンベディングするものである．KL 距離の指標において，すべての他の手法は重み付きランダム (weighted random) より優れている．他の評価指標においても，すべての他の手法は重み付きランダム (weighted random) より同等か高性能である．そういうわけで，エンベディ

表 5.4: 複数ラベル分類の結果．KL 距離は小さいほど良い結果を示し，その他は大きいほどよい結果を示す．

method	KL div.	MRR	MAP
weighted random	40.734e-2	45.000e-2	74.341e-2
Word2Vec	38.810e-2	44.973e-2	74.313e-2
GloVe	**36.187e-2**	47.802e-2	73.608e-2
DeepWalk	39.496e-2	45.192e-2	74.176e-2
LINE(1st)	40.006e-2	45.000e-2	74.341e-2
LINE(2nd)	37.796e-2	51.209e-2	73.553e-2
LINE(concat)	37.560e-2	51.071e-2	73.343e-2
PTE(joint)	39.417e-2	45.192e-2	74.368e-2
proposed(concat $\vec{u_{sc}}$)	40.425e-2	45.000e-2	74.341e-2
proposed(concat $\vec{u_{sr}}$)	38.766e-2	45.082e-2	74.341e-2
proposed(concat $\vec{u_{ss}}$)	38.933e-2	45.137e-2	**74.368e-2**
proposed(divide $\vec{u_{sc}}$)	38.606e-2	45.000e-2	74.341e-2
proposed(divide $\vec{u_{sr}}$)	39.051e-2	45.852e-2	74.167e-2
proposed(divide $\vec{u_{ss}}$)	37.216e-2	**51.511e-2**	72.610e-2

ング法を人流データに適用することは妥当で，それぞれの駅の目的分布を抽出することは妥当で効果的である．

次に，GloVe と他の手法との比較をする．GloVe は KL 距離指標において最も良い結果を示している．というのは GloVe だけがデータセット全体の共起情報を利用しているからだと考えられる．長い文脈の共起情報の効果はまた LINE(1st) と LINE(2nd) の間の結果でも示されている．LINE(1st) は 2 つノード間のエッジの重みを近似するのに対し，LINE(2nd) は 2 つのノード間で共有するノードの近接性を近似する．この効果は KL 距離と MRR の結果にあらわれている．これらの結果は全体のグラフ構造を利用することは複数ラベル分類において良い効果があるということを示している．

表 5.5: パラメータ α の推定結果．

Variable	Average	Std dev.
α	0.468	0.0153

提案モデルを PTE(joint) 法と比較を行う．特に，提案モデルの (divide $\vec{u_{ss}}$) は PTE 法よりも KL 距離と MRR 指標で優れている．さらに提案モデルの (concat $\vec{u_{ss}}$) はま

表 5.6: それぞれの目的ベクトル周辺の 10 駅の地理的な位置の標準偏差.

method	"on business"		"others"		"to work"		"to school"		average	
	long SD	lat SD	long SD	lat SD	long SD	lat SD	long SD	lat SD	long SD	lat SD
PTE(joint)	4.052e-2	6.884e-2	6.283e-2	8.801e-2	4.991e-2	8.412e-2	12.866e-2	9.020e-2	7.048e-2	8.280e-2
proposed(concat $\vec{u_{sc}}$)	6.516e-2	8.739e-2	4.804e-2	7.941e-2	5.530e-2	8.124e-2	6.399e-2	7.714e-2	5.812e-2	8.129e-2
proposed(concat $\vec{u_{sr}}$)	6.126e-2	7.224e-2	7.586e-2	7.782e-2	6.912e-2	7.102e-2	10.655e-2	8.183e-2	7.820e-2	7.573e-2
proposed(concat $\vec{u_{ss}}$)	3.584e-2	7.897e-2	4.025e-2	7.259e-2	3.498e-2	7.174e-2	7.048e-2	7.869e-2	4.539e-2	7.550e-2
proposed(divide $\vec{u_{sc}}$)	10.596e-2	**9.604e-2**	9.551e-2	9.603e-2	9.593e-2	10.133e-2	8.789e-2	8.831e-2	9.632e-2	9.543e-2
proposed(divide $\vec{u_{sr}}$)	11.701e-2	9.416e-2	14.268e-2	11.357e-2	11.731e-2	11.242e-2	14.934e-2	15.349e-2	13.159e-2	11.841e-2
proposed(divide $\vec{u_{ss}}$)	**11.701e-2**	9.416e-2	**14.268e-2**	**11.357e-2**	**11.731e-2**	**11.242e-2**	**14.934e-2**	**15.349e-2**	**13.159e-2**	**11.841e-2**

た KL 距離と MAP 指標で優れている．これらの結果は提案モデルがラベル情報を PTE(joint) 法よりも人流データにおいて効果的に利用していることを示している．

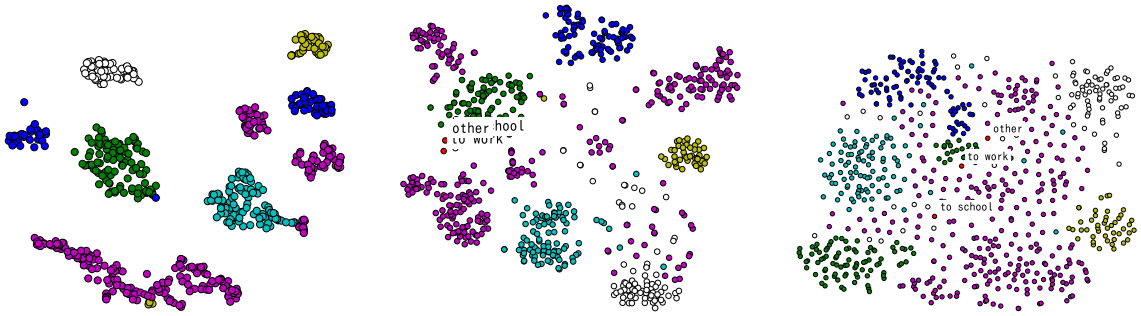
最後に提案モデル間での比較を行う．divide $\vec{u_{ss}}$ は concat $\vec{u_{ss}}$ よりも KL 距離と MRR 指標に関して優れている．この違いは 2 つのモデル間の更新方法の違いから生じている．接続モデルは一度にベクトルの要素の半分を更新するが，内分モデルはベクトルのすべての要素を更新する．この違いは学習したベクトル表現にあらわれている．この違いについて後の節でさらに詳細に述べる．両方のモデルにおいて，$\vec{u_{ss}}$ の結果は $\vec{u_{sc}}$ と $\vec{u_{sr}}$ の結果よりも優れている．これは，地理的な情報と目的の情報の両方が複数ラベル分類に対して必要であることを示している．

パラメータ α の内分モデルの結果を表 5.5 に示す．この結果は $\vec{u_{ss}} = 0.468\vec{u_{sc}} + 0.532\vec{u_{sr}}$ で，これは $\vec{u_{sr}}$ が $\vec{u_{sc}}$ よりも重要であることを示している.

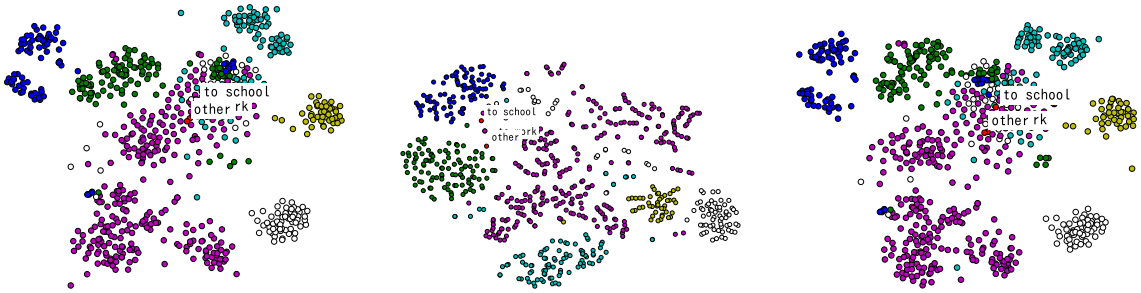
(2) 目的ベクトル周辺の地理的な位置の分布

次に，得られた目的ベクトルの特徴を明らかにする必要がある．そこで，目的ベクトル周辺の駅の調査を行う．本章で述べてきたとおり，駅へ行く目的を正確に抽出することを試みている．そして，駅へ行く目的は地理的な位置とは無関係である．もしそうであるならば，提案手法は目的ベクトル周辺でお互いに地理的な距離は離れている駅同士を集めるはずである．この仮説を駅の地理的な位置の標準偏差を確認することで評価する．結果は表 5.6 に示す．この表はそれぞれの目的ベクトル周辺 10 駅の位置情報の標準偏差である．

divide $\vec{u_{ss}}$ モデルと PTE(joint) 法の比較結果は，divide $\vec{u_{ss}}$ モデルの標準偏差が PTE(joint) 法よりもすべての目的で大きいことを示している．これは，divide $\vec{u_{ss}}$ モデルはそれぞれの目的ノード周辺に離れた駅を集められていることを示していいる．接続 (concat) モデルの結果は他の手法よりも小さい標準偏差を示した．以下の可視化の結果の節で，結果についてより詳細に考察する．



(a) DeepWalk (b) PTE(joint) (c) concatenating model($\vec{u_{ss}}$)



(d) internally dividing model($\vec{u_{ss}}$) (e) internally dividing model($\vec{u_{sc}}$) (f) internally dividing model($\vec{u_{sr}}$)

図 5.3: t-SNE [67] による獲得したベクトルの可視化．それぞれの点は駅や目的を示す．駅は所属する社局によって色分けされている．

(3) 分散表現の可視化

最後にそれぞれの手法の可視化の結果を示す．可視化の結果を図 5.3 に示す． 6 つの手法の可視化の結果を示す．この図において，それぞれの点は駅や目的を表す．そして，この点は所属する 6 社局で色分けされている．図において，(a) と (b) は既存手法で，(c)–(f) は提案手法の可視化の結果である．また，図中でラベルを持つ点はそれぞれ目的ベクトルの推定された座標，ラベルの無い点は駅を表している．そして駅は，その駅が所属する社局によって色分けされている．

(a) DeepWalk の分散表現は会社ごとにクラスタを形成している．この人流データセットにおいて，ほとんどの人々は狭い範囲を通常移動し，あまり乗り換えをしないという統計的な結果を得ている．ゆえに，DeepWalk の可視化の結果は局所的な文脈情報を捉えている点で妥当である．この結果はまた (e) proposed(divide $\vec{u_{sc}}$) でも観察される．線形的に整列している駅が図内で見られるが，これらの駅は同じ路線に沿っている．

(f) proposed(divide $\vec{u_{sr}}$) において，可視化の結果は目的ベクトルの周囲で混交している．さらに (d) proposed(divide $\vec{u_{ss}}$) は (e) proposed(divide $\vec{u_{sc}}$) と (f) pro-

posed(divide $\vec{u_{sr}}$) を合わせたものになる．これによって，提案手法はそれぞれの目的ベクトルの周囲に位置が離れた駅が集まり，表 5.4 における MAP 指標で有用な目的推定の結果を達成している．一方で，図の外側では同一社局の駅でクラスタを形成している．観察の結果これらの駅の大部分は「帰宅」の目的が与えられるべき郊外の駅であった．移動目的を推定するタスクを考えた場合，これらの駅は社局ごとにクラスタになるのではなく，本質的には混交した状態で大きなクラスタを形成すべきであるはずである．同一社局の駅でクラスタ形成してしまう点については,「帰宅」のラベルを除外していること，モデルに利用した二次の近接性では社局を超えて同じ目的で行かれる駅の近接性をうまく学習できないことが考えられる．

5.7 本章での結論

第 5.6.3 節で述べた通り，提案したモデルは PTE 法よりも優れた結果を示した．これらの結果は，地理的な依存性のある大規模移動データにおいて，提案モデルがそれぞれの地域に行く目的の特徴を PTE 法よりもうまく捉えていることを示唆している．人々の移動する領域はたいてい狭く，限られた領域で生活している．この制約に照らすと，提案モデルは PTE 法よりもうまく機能する．複数ラベルの分類のタスクで，提案モデル (concat $\vec{u_{ss}}$ と divide $\vec{u_{ss}}$) は表 5.4 で良い結果を示している．この結果は提案した “移動目的推定モデル” の正しさを強調するものである．特に，ベクトルの可視化の結果 (図 5.3) では，the proposed(divide) モデルはそれぞれの地域を地理的依存のベクトルと目的ベクトルに分解している．最後に，パラメータ α の推定結果は印象的である．この結果は移動する目的が距離よりも 1.1 倍重要であることを意味する．つまり，人々は積極的にやりとげたい目的を持っている場合は遠く離れていても移動する．また，提案手法は移動目的の推定において半教師有り学習の枠組みで少量のラベルデータを元にラベル無しの対象に対するラベルを推定するタスクを改善するものである．ラベルデータが全く無い目的を推定するものでは無いということについては言及しておく．

しかし，現在の提案モデルの性能はわずかに教師なしのエンベディング法より優れているだけである．これは提案手法は単に 2 次の近接性のみをモデル化しているからで，ネットワーク構造全体を捉えていないからだ．次のステップとして，異種ネットワークの持つ全体構造やラベルネットワークのより効率的に学習する方法を考慮すべきである．グラフの全体構造は GraRep 法 [11] や GloVe 法 [51] によってうまく捉えられている．地域へ行く目的を抽出するためにこうした取り組みを参考にする必要があるだろう．

ここで，本研究全体における本章での結論について述べる．本章では，第 3 章で提案した，異種データ間のユーザ移動ネットワークの実データへの適用と知識抽出

の効果の検証を行った．特に本章では，ユーザ移動ネットワークのうち，データソース間でユーザを一意には特定できる場合の検証を行った．データとしては，複数の公共交通の運営会社がそれぞれに管理するユーザの乗降履歴のデータを用いた．このデータは複数の運営会社がそれぞれ管理するものだが，ユーザＩＤは共通となっているため，データソース間で一意にユーザを特定することができる．本章ではこのデータから作ったユーザ移動ネットワークを元に，特にユーザの移動の文脈情報の利用が，知識抽出に効果的であることの検証を行った．ここでは，移動目的推定モデルを提案し，乗降データのような人流ネットワークだけでなく，路線図のような地理的制約ネットワーク，パーソントリップ調査のような目的近接性ネットワークを組み合わせて，知識の抽出を行った．検証の結果，地理的には離れていても同様の目的で移動する地域の分散表現を獲得し，また，地理的な近接性よりも移動する目的の方がより重視されていることがわかった．これによって，第 3 章で提案した，ユーザ移動ネットワークのうち，データソース間でユーザを一意に特定できる場合について，実際のスマートカードの乗降履歴データに適用し，有用な知識を抽出できることを示した．本章での検証によって，ユーザ移動ネットワークを適用することによって，獲得する知識の価値が高まることが示唆されている．

第6章　複数公共データの統合による知識抽出システムの実装と検証

6.1　本章で用いるユーザ移動ネットワーク

第4章，第5章では，第3章で提案した異種データ間のユーザ移動ネットワークに基づいた知識抽出の方法論を述べ，その有効性を示した．しかし，最後にこのユーザ移動ネットワークによる分析が実社会でも有用であるか，議論が残る．そこで，本章では，複数の公共データをユーザ移動ネットワークによる知識抽出システムとして実装し，実務的に日々分析業務を行っている関係者に実際に利用してもらうことで有効性の検証を行う．本章で用いる異種データ間ユーザ移動ネットワークの詳細な図を図6.1に示す．第5章と同様にここでも複数の運営会社によって個別に管理されている乗降履歴に関するデータを用いる．ここで，それぞれの運営会社の分析担当者は，自社の管理するデータの分析は行えるが，他社の管理するデータも含めて分析することはできないと状況がある．そこで，本章では複数データを組み合わせて分析を行えない現状を打開するために，彼らが懸念するリスクを調停し，利害の衝突の無いような分析システムの構築を行う．図に示しているような異種データ間のユーザ移動ネットワークの枠組みを用いることで，このような調停，データの追加，削除を行うことが容易になる．本章では，ユーザ移動ネットワークによる知識抽出を実社会で有効に機能させるために必要な枠組みを示し，評価することを目的としている．

6.2　本章の背景，目的

近年，社会に存在する様々なデバイスのIoT化が進み，あらゆるデバイスが一種のセンサーとして機能し始めている．これによって，人，モノの移動に伴って生じる様々な情報がデータベース上に大規模に蓄積されるようになった．しかし，各企業が持つこうした大規模なデータはそれ自身では活用が難しい面がある．なぜならば，企業に蓄積されるデータは関連する環境や，別の原因によって生じたものであることが多いからである．例えば，鉄道の乗降に関するデータを蓄積していても，ある駅で降車する理由を知ることは難しい．そこで，乗降データに加えて，駅周辺で人々

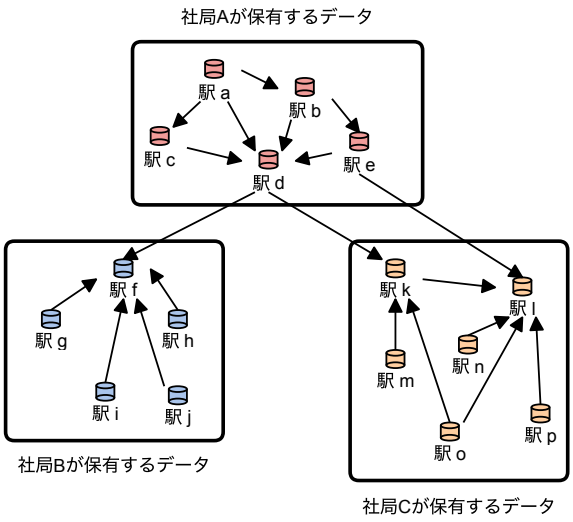


図 6.1: 第 6 章で用いる異種データ間ユーザネットワーク.

が行った経済活動に関するデータ (PoS データ等) があれば，降車理由をある程度類推できるだろう．そして，駅周辺での人々の経済活動に関するデータも PoS データに限らずさまざまな場所，形式で蓄積されている．周辺の他社路線の駅へ乗り換えであったり，駅周辺でのソーシャルメディアのログデータであったりと組み合わせて分析することで有用な知見を得られそうなものは数多く想像できる．

このような大規模なデータの蓄積に伴って，それを活用しようと関連するデータを保有する企業間で提携を行い，お互いにデータを提供することで，人，モノの移動を大規模に捉えようという試みが少しずつなされてきている．しかし，いまだ事例としては少数で，多くの企業ではあまり促進していない．

これまで，異なる情報源から生じるデータの統合に関する研究がなされてきた．代表的なものとして，データ統合に関する理論的背景を与える研究 [66] が考えられる．また，一方で，企業間のデータ提供の試みに関しては,「気象ビッグデータを活用した需要予測精度向上によるサプライチェーンの全体最適化」[88] という事例が報告されている．本章で行う研究では，LaV(Local as View) に基づいたデータ統合スキーマを設計し，それに基づいたデータベース作成と分析システムを提案する．本章では，実際の企業間の提携の際に生じる障害も含めたシステム設計を行い，今後のデータによる企業間提携を促進させる狙いがある．

複数の企業で個別に蓄積しているデータを提供しあい，複合的に分析することは有用である．これに異論を唱える人はいないだろう．しかし，実際にはこの試みはあまり行われていない．なぜならばデータを提供することには障害があるためである．例えば，自社のデータを他社に提供することで，自社の機密情報を知られてしまうのではないか，といった懸念である．また，一方でデータ提供によって便益を

得るのは特定の一社であって，提供した会社自身には何の便益も無いという懸念もある．こうした複数の要因によって，相互にデータを提供し，複合的に分析を行う事例があまり増加していない．

そこで本章では，企業間のデータ提供の促進を図るためにステークホルダー同士のデータ提供を意思決定に役立てるためのフレームワークを提案する．データ企業にはさまざまな事情があり，新たにデータを提供したり，もしくは途中からデータの提供をやめる可能性がある．そこで，データベースを Local as View の方式で疎結合する方式でシステム設計を行った．さらにデータの機密性を守るために第三者の調停者を設置することで，入力クエリや分析結果の検閲を行うようにしている．

提案するフレームワークは，データ統合部と分析システム部の 2 つの機能に分類される．データ統合部では，それぞれの競合企業のデータを調停者 (Mediator) に提供してもらい，ユーザの秘匿化を行う．そして，提携の目的に合わせたレベルにまでデータの抽象度をあげる処理を行い，統合ＤＢの作成を行う．分析システムは複数の企業間を移動したデータを対象にした分析を行うもので，競合企業の機密部分を分析できないように調停者による認証を必要とする．

提案したシステムを関西の鉄道企業 6 社局の提携データを元に実装した．そして，実際の分析担当者に利用してもらい，彼らの持つ自社内でのデータ分析，運営を元に得ている経験知の検証を行ってもらった．結果として，彼らの持つ経験知を社局間を横断するデータベースを用いることで定量的に分析できることがわかった．利用者自身のアンケート結果においても社局間を横断するデータ分析のメリットをすべての担当者が感じていることがわかった．また，自社の機密情報の漏洩についても 71.4 %の担当者が障害とならないことを感じていることがわかった．

本章の貢献を以下にあげる．

1. 競合企業間で相互データ提供に基づく戦略的提携を効率的に解決する分析フレームワークを提案した．
2. データ提携により各企業にとって定量的に便益があることを示した．
3. 調停者による秘匿化，ユーザ認証によって，企業の機密情報の保護を担保したシステムであることを示した．

6.3 関連研究

ここでは，本章で行う研究に関連する研究と本章で提案する手法に関連する研究について説明する．

6.3.1　データ統合と異種データの分析に関する研究

これまで，さまざまなデータの統合に関する研究が数多くなされてきた．代表的なものとして，データ統合に関する理論的背景を与える研究 [66] が考えられる．彼らは GaV(Global as View) と LaV(Local as View) という 2 つの統合方式を提案している．GaV はデータそのものを加工し，統合する方式であるため，この方式を複数の企業が所持するデータに適用する場合，それぞれのデータを完全に共有する必要がある．一方で LaV はデータは共有せず，データへのアクセス方法を共有，統合する方式である．そのため，複数の企業が所持するデータに適用する場合，データそのものを共有する必要がなく，アクセス方法のみを共有すれば良い．そのため，共有元となる企業側で開示するデータの範囲，期間を制限しやすいという利点がある．ただ一方で，GaV はデータを完全に統合できるため，通常クエリに対する応答速度が LaV と比較し高速である，という利点がある．しかし，本章で行う研究では，共有元の企業側の機密情報の保護を重視するため，LaV によるデータ統合スキーマを設計し，それに基づいたデータベース作成と分析システムを提案する．

また，本章で行う研究は異なる種類のデータを統合し，分析を行う試みと考えることができる．こうした試みは実業界でも取り組まれており， 気象情報を利用して，食品ロスを削減しようとする,「気象ビッグデータを活用した需要予測精度向上によるサプライチェーンの全体最適化」[88] という試みが代表的なものとして挙げられる．この試みは食品のサプライチェーン上の様々な企業がデータを提供しあい，気象予測情報を軸として予測した需要を元に材料，流通の最適化を行おうとするものである．この試みは，サプライチェーン上のどの企業にとっても食品ロスという無駄の削減を行うことができるという便益がある．企業がデータ提供を検討する場合，このような便益を受ける見込みが低ければ，データ提供に伴う企業機密の漏洩のリスクを恐れて，提供を断念する懸念がある．そこで，本章で行う研究では，システムの枠組みに調停者を導入することで，便益の担保と懸念の低減を行うこととした．

6.3.2　ステークホルダーと戦略的提携に関する研究

ステークホルダーに関する研究は， Freeman によって始めれたと言って良いだろう．ステークホルダーという定義は様々な文献で多種多様に定義されており，範囲が明確でない．本章では，Freeman が著書の中で定義した「ステークホルダーは，組織体の目的の遂行に影響するか影響を受けるグループまたは個人である」[18] という定義を採用する．したがって本章でさすステークホルダーは，株主，従業員，顧客だけを指すのではなく，競合関係にある企業，メディア等も含まれる．そしてこれら複数のステークホルダー間でデータを共有し，意思決定支援を行うシステムについて議論を行う．

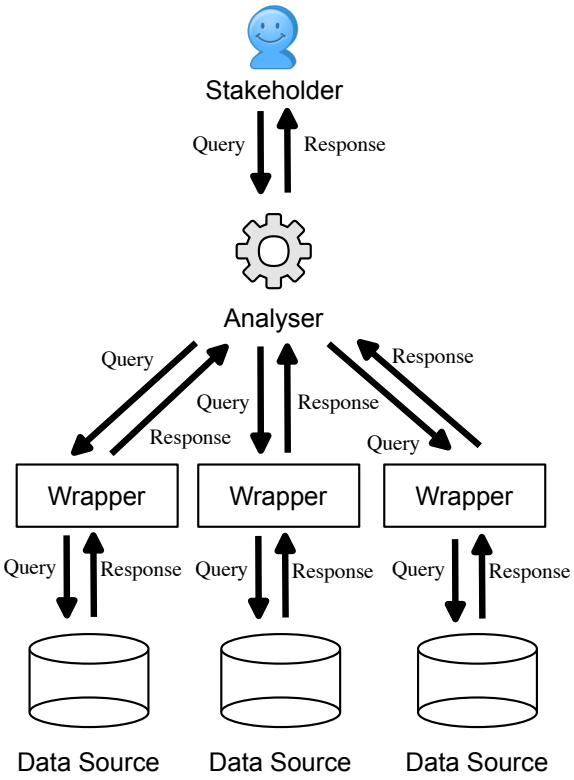


図 6.2: 提案するフレームワーク.

一方で，それぞれの企業間でデータの提供を行う行為は，相手方の知識を移転させる [46] という意味において企業間の戦略的提携の 1 つと捉えることも可能であろう．戦略的提携は，企業間の部分的結合を伴う場合 (資本提携) と伴わない場合 (契約提携) に大別できる [89]．戦略的提携は，企業同士で共通の目標を達成するために，知識や技術の移転のために行われることが多い．しかし，単にデータを相互に共有する場合に，企業間で共通する目標を持つことは難しいことが多い．なぜならばそれぞれの企業の置かれる立場，経営環境は異なっており，提供されたデータをどう利用するかは各企業が決定するためである．そこで，本章ではこのデータ提携という企業間の結びつきが比較的弱い契約提携の中でも，結びつきの弱いデータ提供を元にそれぞれの企業がそれぞれの目的を達成できるような意思決定支援システムの構築を目指す．これによって，共通の目的意識を持ちにくい企業間においてもデータ提供によるメリットがわかりやすくなり，データ提供による企業間提携がより一層促進されるだろう．

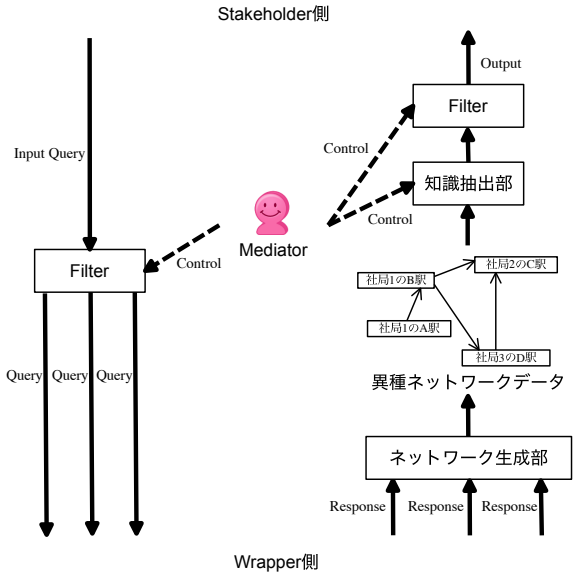


図 6.3: 分析部分の詳細.

6.4 マルチステークホルダーによる相互データ提供に基づいた意思決定支援のフレームワーク

本節では提案するフレームワークとその詳細について説明する.

6.4.1 概要

提案するフレームワークの概要を図 6.2 に示す．このフレームワークは 3 つの部分から成り立っている．すなわち，ステークホルダー，分析部，ラッパーとデータソースの部分である．まず，図中のステークホルダーは複数のステークホルダーのうち，分析を行おうとしている人（または法人）を表している．次に，データソースは様々なステークホルダーから提供されているものである．このデータソースは，提供者の都合によって，新たに提供されたり，提供されなくなったりする．そのため，これらのデータソースを一体化したデータベースを作る (Global as View approach: 以下，GaV とする) のではなく，個別のデータソースへアクセスするためのクエリに変換するラッパーを経由して利用する (Local as View approach: 以下，LaV とする) データベースを利用する．そして，それぞれのデータからの返信を分析部への入力とする．分析部の詳細については次節で説明を行う.

このフレームワークにおいて，分析を行うステークホルダー自身が分析するための問い合わせをする．そして分析部で，この問い合わせを解釈し，各データソースに向けてクエリを発行する．各データソースに向けたクエリはそのデータソースに

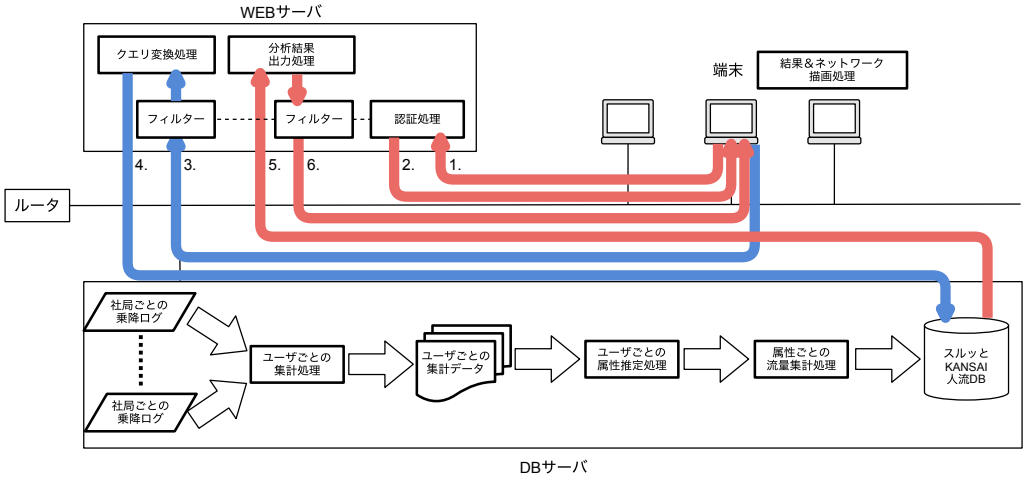


図 6.4: 実装したシステムの概要.

向けた形式にラッパー部分で解釈し直される．各データソースではクエリに対する返信を送信する．返信内容はラッパーで変換され分析部で統合，解釈され，ステークホルダーの問い合わせに応答する内容を返信する．

6.4.2　分析部分の詳細と調停者の役割

分析部の詳細を図 6.3 に示す．分析部分は 4 つの部分から成り立っている．すなわち，調停者，ネットワーク生成部，知識抽出部，フィルターの部分である．まず，ネットワーク生成部は，提供されたデータソース群から分析対象となるデータを抽出し合成する機能を担っている．次に知識抽出部は，入力されるネットワークデータに対し，分析の主体となるステークホルダーの目的にあった手法を用いて，知識抽出を行う．フィルターは主に分析の主体以外のステークホルダーにとって損失となるようなデータ漏洩や分析が行われないようにするものである．最後に，調停者は本フレームワークにおいて大きな役割を担う．調停者は人とは特定せず，電子的な制御装置を用いても構わない．調停者は，主に他のステークホルダーと主体となるステークホルダーの利害関係を調整する役割と，主体者の目的と合致した知識抽出手法を選定する役割を担っている．

6.5　試作システムによる検証

本章では，6.4 章で提案したフレームをシステム上に試作し，検証を行う．

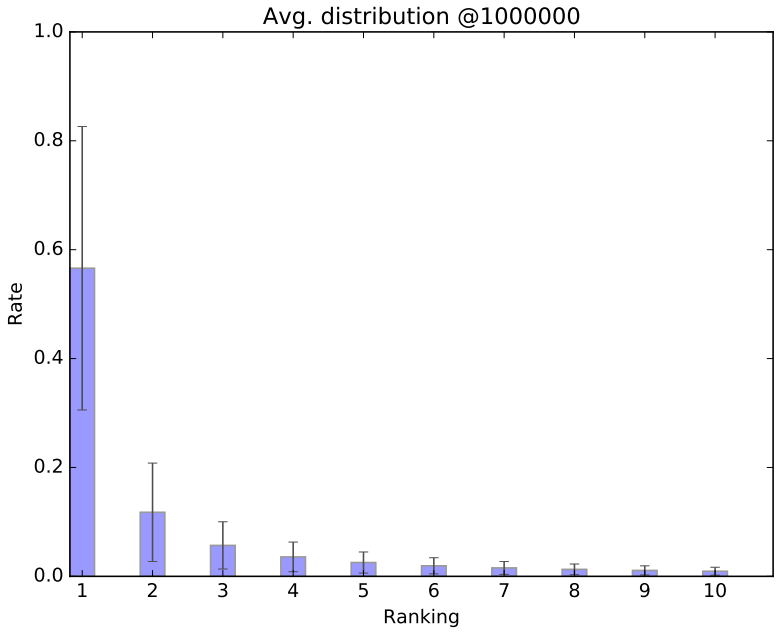


図 6.5: ランダムに選択した日常的に利用するユーザ 100 万人の往復の経路の分布．

6.5.1　使用するデータ

使用するデータは関西圏の鉄道事業者 6 社局から提供された，非接触型ＩＣカードユーザの乗降ログを利用する．本データは事業者側でデータの匿名化化を行っており，復元できない形になっている．データの内容は，各ユーザの性別，年代と乗降日時，乗降駅となっている．今回は 2013 年 04 月 01 日〜2016 年 05 月 31 日の乗降データを利用した．

6.5.2　提案フレームワークの具現化

この鉄道事業者から提供されたデータを元に，6.4 章で提案したフレームワークに則って分析システムの試作を行った．試作したシステムの構成を図 6.4 に示す．このシステムは主に 3 つの部分から成り立っている．それは，WEB サーバ，DB サーバ，端末である．

端末は，ステークホルダーの分析を行う担当者が利用し，分析するための問い合わせを送信したり，受信した内容の表示，ネットワーク描画処理を行う．

WEB サーバは，４つの機能を担当している．それは，認証機能，フィルター，クエリ変換処理，分析結果出力処理である．認証機能は端末を操作している担当者を認証する．担当者の認証は，クエリの入力や出力結果のフィルター機能と連携して

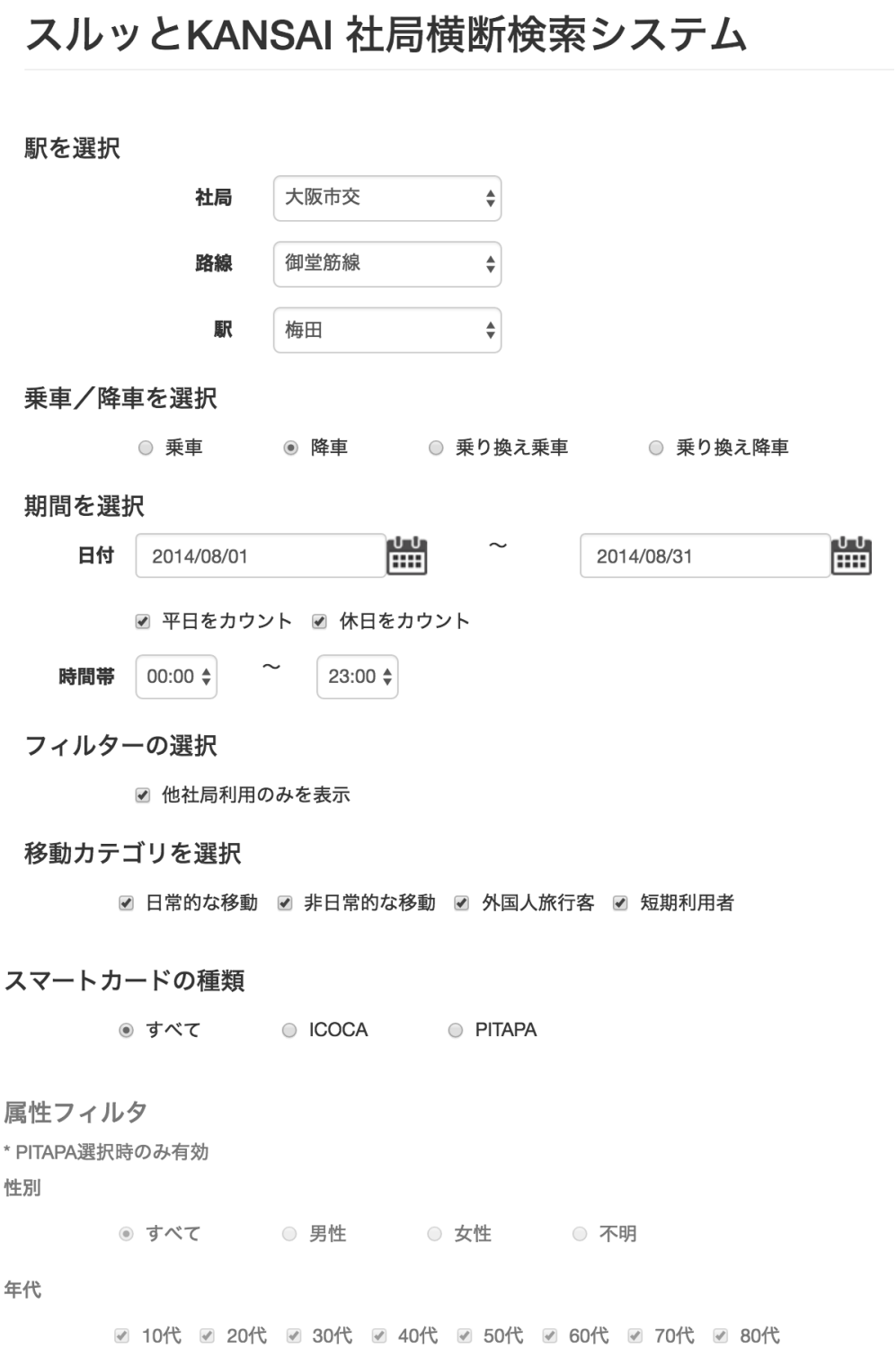


図 6.6: 試作したシステム（検索画面）.

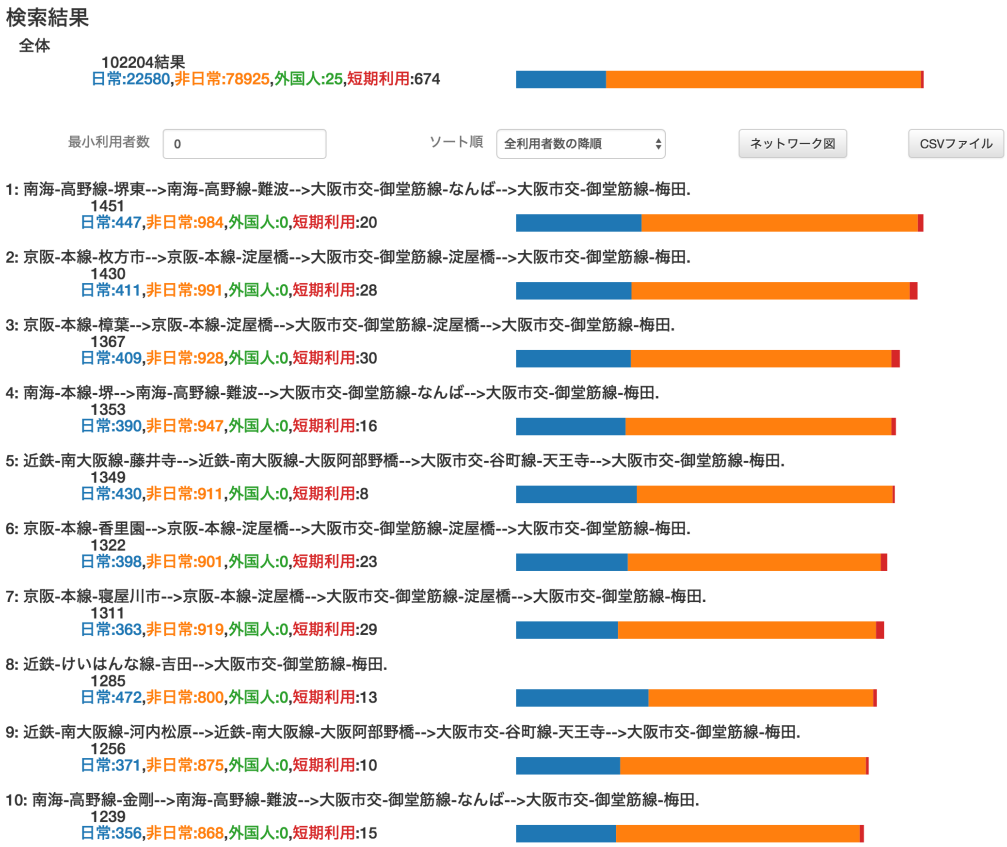


図 6.7: 試作したシステム（結果画面）.

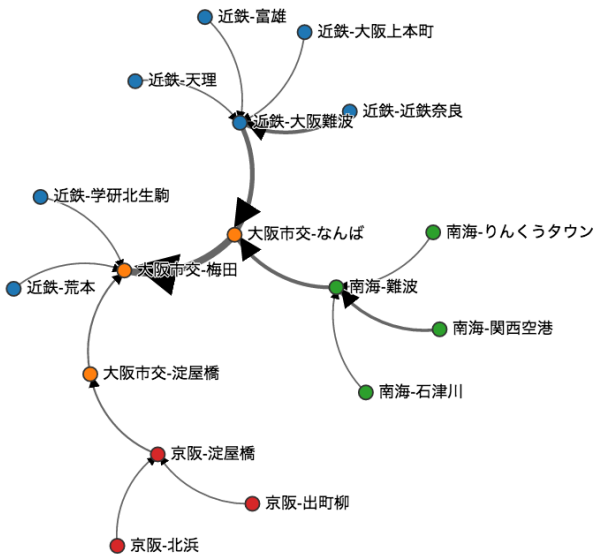


図 6.8: 試作したシステム（ネットワーク描画画面）.

おり，他のステークホルダーとの利害の衝突の回避を行う調停者の機能も担っている．クエリ変換処理は，端末から送信されてきたクエリを DB へ問い合わせるための具体的なクエリへ変換する．分析結果出力処理は，DB サーバから送信されてきた返信内容を端末で表示するのに適した形式へ変換する．

DB サーバは，複数の鉄道事業者から提供されたデータを元に集計した人の移動に関する DB(スルッと KANSAI 人流 DB) を保持している．

6.5.3　分析機能と調停者機能の実装

提案したフレームワークでは，分析機能はステークホルダーのクエリに合わせて調停者が選択することになっている．今回のシステム試作にあたり，各ステークホルダーに個別にヒアリングを行い，共通に関心のある分析内容の選定をした．その結果，自社局が運営する路線内の駅での乗降客と他社局の路線との関係性について，すべての社局が興味を持っていることが判明した．そこで，筆者らが調停者となり，分析機能として自社局内の利用者の他社局内での利用状況を検索できる機能を実装することとした．また，フィルター機能として，他社局路線内の利用状況の分析結果は他社への秘匿情報となるため，認証機能によって，他社局の運営する路線情報は選択，表示をできないように実装を行った．

さらに詳細なヒアリングの結果，特に外国人観光客，日常的に利用するユーザの動向に関心が強いことが判明したので，国際空港のある駅で初めて／最後の利用があった短期利用者を外国人旅行客と判定し，また，長期利用者が 40%以上の割合で利用している往復の経路を日常的な利用経路として判定行い，ユーザごとに推定を行い，結果としてそれぞれの推定結果も合わせて表示するようにした．日常的に利用するユーザをランダムに 100 万人選択し，それぞれのユーザの往復の経路の分布を調べたものを図 6.5 に示す．図に示した通り，多くのユーザは主となる往復経路を日常的に利用している．約 66%のユーザが主となる往復経路となる 40%をしきい値として選定した．

6.5.4　試作システムによる検証

試作システムを用いて，提案したフレームワークの効果について検証を行った．分析主体となるステークホルダーは，自社内のサービスを改善したい鉄道会社である．

実装したシステムの端末上での検索条件の入力画面を図 6.6 に示す．ヒアリングの結果を踏まえて，乗客の他社局との行き来の情報，ユーザの属性の分布を期間，時間帯を区切って検索できるように実装している．この検索画面で条件を設定し，クエリをサーバへ送信すると，検索結果が図 6.7 のように表示される．この結果は，大

阪市営地下鉄・御堂筋線・梅田駅で 2014 年 08 月に降車した人々の他社局とからの流入状況の分布を表したものである．

本システムでは，駅の流入／流出状況を一覧として見やすくするためにネットワーク描画機能を実装している．描画例を図 6.8 に示す．この例は，外国人旅行客の梅田への流入状況を示したもので，エッジの線が太いほど多くの外国人旅行客がその駅間を移動していることを示している．このネットワーク図から大阪の梅田へ，関西空港駅，奈良，京都方面（出町柳）から流入していることが一覧できるだろう．

6.6　検証結果

本章では，検証結果について説明する．6.5 章で述べたシステムを実際に分析を行う担当者に使用してもらい，アンケート結果を取得した．アンケート取得方法は以下のとおりである．アンケートに回答した担当者は，関西の鉄道会社 6 社局の担当者で，2016 年 09 月 05 日に全員に集まってもらい，約 2 時間試作したシステムを使い，分析をするワークショップを行った．ワークショップ前にシステムの使用方法についての説明を 30 分行い，全員に使用方法を理解してもらった上で使用してもらっている．ワークショップを行ったあと，アンケート用紙を配布し，個別に記入をしてもらった．本章では、この取得したアンケートについて説明を行い，その集計結果，各分析担当者によって行われた分析の結果について説明を行う．この，ワークショップ，アンケートに関する詳細は，第 8.3 章（付録Ｂ）にて説明を行っている．

6.6.1　分析担当者へのアンケートの内容

各担当者へ質問したアンケート内容は大きくわけて以下の３つである．

1. 日常的にデータを用いた分析を行っているか．
2. 他社データを含めた分析に利用価値を感じるか．
3. 試作システムに有効性を感じるか．
4. 自社の機密は守られていると感じるか．

このアンケート内で，質問 2,3 に関しては５段階で回答してもらった．5 段階は，5:非常に価値が／有効である，4:十分に価値が／有効である，3:価値が／有効である，2:あまり価値が／有効でない，1:全く価値が／有効で無い，とした．よって，3 以上の回答の場合，質問に対して肯定的な意見となる．また，質問 1,4 は「はい」,「いいえ」の選択方式とした．

ワークショップには 6 社局の担当者計 14 名が参加した．アンケートへの回答は 7 であった．

6.6.2　分析担当者へのアンケート結果

分析担当者へのアンケート結果は以下の通りである．まず，質問 1 の普段からデータを用いた分析を行っているかについては，6/7 が「はい」という回答であった．よってアンケートの回答の大半は社内で普段から分析，レポート業務を行っている人の意見である．

次に質問 2 の他社データを含めた分析に利用価値を感じるか，という質問について，まずすべての回答が 3 以上の価値があるという回答であった．平均では，3.86 であった．これらの結果から自社内のデータだけでなく，他社局のデータを利用できるようになることに大きなメリットを感じていることがわかった．特に担当者からは「従前のアンケートに基づいたデータと比較し，客観性のある乗降客数のデータは信頼性が高く，非常に価値を感じる」という感想があった．

質問 3 の試作システムの有効性に関する質問について，すべての回答が 3 以上の有効であるという回答であった．平均では，3.57 であった．今回は，システム設計者が調停者として，個別に担当者からヒアリングを行い,「駅での乗降者数と経路」に関する分析を行うシステムを設計した．この結果から調停者としての役割が有効に働いていることがわかった．また，一方で「出力した情報の加工が困難」という意見があり，試作システムの機能不足も指摘されている．

最後に質問 4 の自社の機密が保持されていると感じるかどうかについては，5/7 が保持されているという回答であった.「いいえ」と答えた担当者の意見を確認すると，情報の開示範囲については相互に合意を得ておく必要があるという認識であった．今回は試作システムであるので，社局間で公式な合意を結んでいないため，このような回答となったことがわかった．よって，調停者が適切に機能すれば機密は十分に保持されていることが確認された．

6.6.3　分析担当者による分析例と結果

ここではワークショップ時に行われた社局担当者によって行われた分析結果について説明する．特に他社データも含めて分析することでわかった新たな知見について注目して説明する．

(1)　天神祭に来る人々の居住地分布

これは大阪で最も人気のある催事の一つである天神祭に集まる人々がどの地域から来るのかを分析したものである．分析結果として，大阪府内から来る人が多い一方で京都府内から来る人が少ないことがわかった．これは，担当者の認識と一致しているが，本システムによって客観的な数値として示すことが可能となった．また，新たに和歌山から来る人が担当者の認識と比較して多いことがわかった．

これらの結果はすべて他社局のデータを含めることで初めてわかる内容である．

(2)　外国人観光客の動向

複数の社局担当者が大国人観光客の動向について分析を行っていた．多くの担当者が外国人観光客の移動について非常に興味を持っていることがわかる．1 年ごとに一定期間の外国人旅行客の旅客者数を調査し，奈良を訪問する外国人旅行客が増加していること，堺市内に宿泊していると思われる外国人旅行者が増加していること，花見の時期に花見の名所を訪問する外国人が増加していること，等がわかった．これらの結果も同様に他社局のデータを含めることで初めてわかる内容である．

6.7　考察

第 6.6 節で示した結果について，考察を行う．

6.7.1　各社局の分析担当者に行ったアンケート結果について

本システムは日常的にデータ分析を行っている方，鉄道の運営に携わっている方向けのシステムである．その点で，アンケート回答者の全員が評価を行う対象者として適切であった．また，今回他社局のデータを確認できるようにしたことに関しても，すべての評価者にとって有用であることが認められている．特に，評価者の一人からは「従前からある公的データは，利用者の恣意的なアンケートと拡大係数によるものであり，本データは比較にならないほど信頼性が高く，利用価値があると考えます.」という意見を頂いた．これまで，鉄道各社は他社局での自社の利用客の統計データを国内で数年に一回国家的に行われているパーソントリップ調査という大規模なアンケート調査に依拠していた．今回試作したシステムによってこの調査の一部を代替し，しかも日常的に確認できるようになる点を評価していることがわかる意見である．試作システムの有効性については，データの価値と比較すると平均点は低いが，有効であると認められている．評価者からの意見によると，試作

システムでは分析結果の加工，集計が煩雑であるというものがあった．業務効率の向上という観点からの意見だと思われ，実運用上では重要となるものと考えられる．最後に自社の機密に関する質問結果では，一部の社局から担保できていないという指摘をされた．前章で説明したように指摘内容は開示範囲に関する議論のものであった．今回，調停者機能をシステム内に組み込んだが，合意が十分に取れていなかったことが原因と考えられる．今後，各社局の担当者と調停者を含めた形で対話を行い合意を形成することで解決するだろう．

6.7.2　各社局の分析担当者が分析した事例について

本章では，各社局の分析担当者がワークショップで行った分析事例について，一部を示した．数値の詳細については，各社局の経営機密に触れるため公開を控える．しかし，彼らの検証した仮説は，すべて自社線に関する分析であった．仮説そのものも実際に日々鉄道運営に携わっている方ならではの興味深いものであったのと同時に，検証結果も複数の社局に渡るデータを分析することで初めてわかる知見であった．特にさまざまなステークホルダーから提供されたデータを元に分析した場合，特定のステークホルダーにのみ便益があるものになりがちである．しかし，本システムは，調停者がシステム設計前にヒアリングを行うことで，すべてのステークホルダーに便益が得られるように配慮を行っている．データを提供することでそのステークホルダーにどのような便益を得ることができるか，もしくはそのステークホルダーが行いたい分析を可能とすることができるのか，という事前の検討とその他のステークホルダーの要請との調整が重要であることがわかる．本章で検証したシステムによって，提案したフレームワークの有効性が確認された．

6.8　本章での結論

本章では，複数公共データの統合による知識抽出システムの実装と検証を行った．提案したフレームワークに基づいた試作システムによって，以下のことが検証された．まず，相互データ提供時に起こりやすい，便益の不均衡を解消できることがわかった．また，それによって，各企業にとって定量的な便益を得られることがわかった．最後に調停者が適切に機能することで，提携社間での情報の開示範囲，機密の保護を担保できることがわかった．以上のことから提案したシステムが有効に機能し，企業間のデータ提携がより促進されることが期待できる．

ここで，本研究全体における本章での結論について述べる．本章では，第 3 章で提案した異種データ間のユーザ移動ネットワークと，第 4 章，第 5 章で検証したそれを用いた知識抽出の効果について，実社会での課題解決と実用性についての検証を

複数の公共交通の運営会社が管理する乗降データを用いて行った．検証の結果，各企業の分析担当者から，調停者機能が有効に機能することで機密漏洩のリスクが減少していること，ユーザ移動ネットワークを用いた分析に実務上の有効性があることがわかった．これによって，ユーザ移動ネットワークを適用することによって，獲得できる知見の価値が高まると同時に，データを相互に提供するリスクも低減できることを示した．

第7章　考察

本論文では，異種データ間のユーザ移動ネットワーク基づいた複数公共データの統合による知識抽出手法を提案し，その検証のために 2 つの実データを用いた有効性の検証と実社会での効果に関する検証の 2 つの側面から分析を行った．ここでは，これらの分析から得られた知見を元にユーザ移動ネットワークの効果についての考察を行う．

7.1　ユーザ移動ネットワークの効果に関する考察

本論では，第 5 章で鉄道の乗降ログデータを用い，ユーザを異なるデータソース間で同定できる場合の分析を行った．特に第 5 章ではユーザ行動の分析に文脈情報を用いた手法を適用している．これまでの研究では，ユーザの行動が取得できる場合にはそれぞれの行動を自然言語処理や情報検索の分野で伝統的に用いられてきた Bag-of-Words[22] モデルに類する手法を用いてきた．例えば一緒に購買される商品の組を発見するバスケット分析においてはアプリオリアルゴリズム [1] に代表されるようなアソシエーションルールを学習する方法が用いられる．この手法では，ユーザが同時に購入した商品の組を入力として用いる点で Bag-of-Words モデルと類似している．また，文脈情報の利用方法の 1 つである Skip-Gram 法が SPPMI と呼ばれる相互情報量の一種の行列分解と同等であることが Levy らによって示されている [40]．相互情報量は，2 つの語の共起頻度を元に算出されるものであり，Bag-of-Words モデルを発展させたものと考えることができる．つまり，文脈情報を利用することとは，Bag-of-Words モデルを適用する範囲を変化させつつ適用することに等しい．特に，第 5 章においては，駅同士の分散表現上での近接性をユーザ行動を通して，計測しようとしている．これによって，近接する駅間に共通する性質，すなわち移動する目的を見出すことを可能にしている．本章での貢献によって，ユーザ移動ネットワークを用いることで，ユーザ行動の意図を理解することに有効であることを意味している．以上のことから，今後これまでの Bag-of-Words モデルに類するものを入力として行ってきた分析モデルに積極的に文脈情報を適用する方法が発展していくだろう．

一方で，第 4 章で複数のメディアでの露出と検索量のデータを用い，ユーザを異な

るデータソース間で同定できない場合の分析を行った．ここでは，メディアでユーザが接触し，興味を持てば検索を行う，という仮定をおいている．これは複数のメディア間でユーザの同定ができないため設定した仮説である．このように異なるデータソース間でユーザの同定ができない場合には何らかの仮説を設定する必要がある．ユーザ移動ネットワークを作る際に実際にはユーザの移動が無いにも関わらず，ユーザ移動があると仮定してしまうと間違った分析結果を得る可能性があり，擬似相関や交絡といった可能性にも注意を払うべきと考えられる．

最後に実社会での効果について，第 6 章で検証を行った．実務担当者の感想で印象的であったのは,「自社路線に来るお客様がどこから来てどこに行くのか初めてわかった」というものであった．これは異種データからユーザ移動ネットワークを形成する価値について端的にあらわしている言葉であろう．すなわち，単一のデータソースからの情報ではその分析の客観的な価値，位置づけがわからないという問題を表している．この点において，ユーザ移動ネットワークは実社会においても大きな効果を有していると言える．

以上のことから，異種データ間のユーザ移動ネットワークの有効性について，実際の分析事例と実務担当者による評価の両面から有効性が示された．

7.2　ユーザ移動ネットワークと情報統合との関係性

本論文で提案した異種データ間のユーザ移動ネットワークと情報統合との関係について考察する．情報統合とは,「異なる情報源間にある情報を結合し，ユーザに一元化されたビューを提供すること」である [39]．提案したユーザ移動ネットワークもデータソース同士をユーザの移動で結合し，一体として分析を行う，という点で情報統合の 1 つとして捉えることができる．この点で，本節では情報統合との関係について議論する．

まず，情報の結合は情報の保管場所を物理的に一体とする必要はなく，ユーザからは一体となっているように見えていれば良い．そのため，情報そのものはそれぞれのソースに保持したままで，ビューで発行されたクエリに基づいて，該当する情報のみを出力する Local as View(LaV) という統合方法もある．この LaV による統合方式は，データの保存形式と出力するデータ形式を一致させる必要がなく，その形式は各データソースの保有者によって自由に設定できる．ビューへのクエリを各データソースに向けたクエリへどう変換するかという部分のみを設計すれば良い．このしくみによって，データソース内のデータはその保有者が管理すればよく，その開示の範囲や制限を自由に設定することが可能になる．特に事業者にとって，自社が保有するデータソースの情報漏洩とユーザプライバシーの保護は重要な課題である．この LaV のしくみは異なる事業者がそれぞれに管理する高い秘匿性の必要な

データソースからの情報を結合し，利用する場合には必須の機能である．

情報統合という観点で捉えた場合，ユーザ移動ネットワークはデータソース間のユーザの移動によってデータソース間の結合を行う．このため，データソースの管理者にとってユーザのプライバシーの保護に敏感になることが懸念される．本論では，ユーザの特定が完全に可能な場合と部分的にできる場合との 2 つの場合によって検証を行い，特定が部分的にしかできない場合には何らかのモデルによる結合が必要であることを示している．もちろん，ユーザの移動を一意に特定できる方が詳細な分析が可能である．Kevin Kelly も著書 “The Inevitable” において，ユーザ行動のトラッキングとその活用は今後不可避なものと述べている [29] 通り，今後はユーザをデータソース間で一意に特定した形でのユーザ移動ネットワークを利用して分析されていく傾向になるだろう．

ユーザ行動がトラッキングできる状態で想定されるシステムの一例は第 6 章で示した．ここでは，情報を管理している各事業者にとっての機密と事業者間で共有して構わないものについての境界について議論を行い，他社を経由してきた自社利用のユーザの情報を経由元の事業者に提供することにした．各事業者にとって，自社の営業実態を完全に共有されるのは大きなリスクだが，他社での自社顧客の行動には大きな関心がある．このため，他社も併用しているユーザに関してはそれぞれの利用実態を共有することになった．このように，各事業者のニーズと懸念のバランスを取り，それぞれの会社にとって有益な分析結果を得られるような環境を整備することがユーザ移動ネットワークを利用した分析を促進させるのに重要である．

第8章　おわりに

本論文では，各公的機関，事業者が持ちうるデータを統合して知識を抽出するために必要なフレームワークとしてユーザ移動ネットワークを提案した．ユーザの情報が異種データ間で部分的に特定可能な場合と完全に特定可能な場合の 2 種類のデータによりフレームワークの有効性を示し，実証実験にてユーザを介した複数公共データの統合が知識抽出に有効に機能することを示した．

まず，第 4 章では，ユーザの特定が部分的に可能な場合としてマスメディアとソーシャルメディアデータから検索量の推定を行った．ここでは，各々のデータの背後にあるユーザの行動を仮定することでユーザ行動をユーザ移動ネットワーク上のデータ間の関係性から推定できることを明らかにした．本章において，ユーザ行動に仮定をおいたモデル（具体的には，AISAS モデルと SIR モデル）を用いることで分析が可能になることを示した．

また，ユーザを一意に特定可能な場合として第 5 章でスマートカードの乗降データを用いた地域へ向かう目的の推定を行った．特に移動目的推定モデルという，ユーザ移動の系列情報 (文脈情報)，地域間の地理的な近接性の情報，少量の移動目的データを用いることで，地域へ向かう目的を推定する手法を提案した．この手法によって，地理的な移動コスト（距離）の制約を除去して地域へ向かう目的推定を行うことが可能となった．乗降データへ提案手法を適用することで，ユーザの移動の系列情報 (文脈情報) を利用することで地域へ向かう目的を効果的に推定できることを示した．

さらに第 6 章では，異種データを管理する事業者間がデータによる連携をとるためのシステムを提案し，実証的に検証することで事業者間の連携における考察を行った．この検証では，実社会での事業者間データ提携のためにユーザ移動ネットワークを検索，表示するシステムを開発した．そして，鉄道運営会社の実務担当者が実際に使用・評価することで提案したシステムの有効性を示した．特に自社のデータだけでなく他社のデータも利用することで有効性が飛躍的に高まるという点で高く評価されており，複数のデータを組み合わせて分析を行うことで得られる価値の高さを示していると言えるだろう．また，各社間の便益・不利益の調整のための調停を行うことが重要であることも示した．

以上のことから，第 1 章で提起した,「複数公共データの統合・活用に関する技術

的な課題及びデータ保有者間の協力に関する障壁が存在していることにより，複数のデータを分析することで得られる知見の価値が，提供するリスクと比較し低い」という課題の解決のために，提案した異種データ間のユーザ移動ネットワークを適用することが有効であることが示された．具体的には，データ間でのユーザの特定の程度に応じて，分析方法を切り替えることで得られる知見の価値が高められることが示された．それは，ユーザが特定可能な場合には文脈情報を用いた分散表現の獲得手法が有効で，部分的に特定可能な場合にはデータ間の関係性を推定することが重要である．また社会実装においても，ユーザ移動ネットワークを用いることで，データ保有者の利害調整や便益の不均衡を調整でき，データ提供のリスク低減に有効であることが示された．しかし一方で，定量的な効果については，分析事例が少数であることから，他のデータを用いた分析や活用事例を用いてさらに詳細に検証することが必要だと考えられる．

ここで，データの相互提供による事業者間提携の促進という観点からの結論を述べる．冒頭で述べたとおり，データ相互提供の進まない原因として，提供によって得られる知見の価値が低く，それに比して提供するリスクが高くなっていることを障害としてあげた．第 4 章〜第 6 章では，本研究で提案した異種データ間のユーザ移動ネットワークを利用することによって，知見の価値を高めることが可能となり，調停者を正しく機能させることで，リスクも減少させることができることを示した．第 6 章でワークショップに参加した担当者からも複数データを用いた分析によって，予想よりも価値の高い知見を得られた，という感想をいただいている．本ワークショップの成果を受けて，今後参加企業間での相互データ提供を積極的にすすめて行くことが期待されている．以上のことからユーザ移動ネットワークを適切に利用することによって，データの相互提供がこれまでよりも促進される可能性が高まると言えるだろう．

今後，IoT の普及，公的機関，事業者間のデータ提携が広まるに連れて，ユーザをデータ間で特定可能な場合が増加すると見込まれる．すでにウェブ上では，ユーザ行動のトラッキングデータは広告提示や EC 上の商品推薦など，様々なシーンで有効であることが示され，広く普及している．また，国内では T-point カード [1] のような取り組みで，様々な店舗での 1 人のユーザの消費状況をモニタリングし，マーケティングやサプライチェーンの最適化に役立てようと試みられている．今後はウェブ上で行われているような個人化された行動予測が，実世界でもより積極的に行われるようになっていくだろう．つまり，第 5 章で提案したような文脈情報を行動予測に利用する方法は重要性を増していくと考えられる．

また，将来，ありとあらゆるデバイスで取得されるデータがユーザと一意にひもづけられた社会が実現された場合であっても本研究で提案したユーザ移動ネットワー

[1] http://tsite.jp/

クの有用性はあると考えられる．なぜならば，必ずしもユーザとは直接ひもづけられないデータというのは様々に考えられるからである．例えば，ユーザがときどき出入りする，密閉された小さな個室の室温のデータを予測する必要があるとしよう．この室温は，個室のデータであるので，直接ユーザにひもづけられるデータではない．しかし，人の出入りによってこの個室の室温が影響を受けることが明らかな場合，本研究で提案したユーザ移動ネットワークのフレームワークは適用することができる．そして，ユーザの出入りと室温データの関係性を検証することが可能になる．つまり，将来にわたって本研究のフレームワークは有効であると結論付けられる．

ここまで，ユーザ移動ネットワークを用いることで，データ相互提供が進み，将来的にはユーザをデータ間で一意に特定されるようになっていくことを説明してきた．最後にそれによって実現される社会について議論する．本稿では，これまでユーザ移動ネットワークをさまざまな分析に適用し，議論してきた．第 4 章では，マーケティング担当者が商品ＰＲへの施策を考える上で有用であることを示した．第 5 章では，駅への移動目的を推定することで，都市，沿線の開発，ＰＲに有用であること示した．第 6 章では，各事業者の分析担当者からみて，運行上のサービス改善に有用であることを示した．このようにユーザ移動ネットワークによる相互データ提供が進むことで，データを保有している事業者，または分析担当者の人々に便益をもたらす．分析に得られた知見の社会実装が進めば，対象のサービスが改善されることが見込まれ，事業者だけでなくデータを提供している公共機関，事業者のサービスを利用しているユーザ体験の改善が見込まれるだろう．

付録 A:第5.4節の更新式の導出

8.1 LINE(2nd) モデル

本章では，ネットワークエンベディングモデルとして，“LINE(2nd) モデル” [64] を採用している．このモデルの目的関数は，5.2 式であるが，この式の詳細な導出から始めたい．

まず目的関数を，式 8.1 のように定める．ここで，$\hat{p_2}(\cdot|v_i)$ は観測できる頂点 v_i からのエッジの重みの分布，λ_i は頂点 v_i の重み，$d(\mathbf{x}, \mathbf{y})$ は 2 つの分布の距離を表すものとする．

$$O = \sum_{i \in V} \lambda_i d(\hat{p_2}(\cdot|v_i), p_2(\cdot|v_i)) \tag{8.1}$$

簡易的に λ_i を頂点 v_i の次数，$d(\boldsymbol{x},\boldsymbol{y})$ を KL 情報量，$\hat{p_2}(v_j|v_i) = \frac{w_{ij}}{degree_{v_i}}$ と定めると，この式は定数項を削除し，式 8.2 で表される．

$$O = -\sum_{(i,j) \in E} w_{ij} \ln p(v_j|v_i) \tag{8.2}$$

本稿では SGD 法によって最適化を行うが，$p(v_j|v_i)$ の計算には，頂点 v_i から張られるすべてのエッジについて計算が必要なため，計算量が必要となる．そこで，ネガティブサンプリング法によって概算を行う．

$$\ln p(v_j|v_i) = \ln \sigma_{j'i} + \sum_{k=1}^{K} E_{v_n \sim P_n(v)}[\ln \sigma_{-n'i}] \tag{8.3}$$

ここで，$\sigma_{j'i}$ は，$\sigma(\vec{u_j'} \cdot \vec{u_i})$ を，また，$\vec{u_j'}$ は，頂点 v_i から遷移する先の頂点 v_j を表すベクトルで，文脈 (context) ベクトルと呼ばれるものである．

これらの 8.1 式〜8.3 式を用いて，各目的関数の最適化において実際に用いる更新式を導出する．

$$\frac{\partial O_{ij}}{\partial \vec{u'_j}} = -(1-\sigma_{j'i})\cdot\vec{u_i} \tag{8.4}$$

$$\frac{\partial O_{ij}}{\partial \vec{u'_k}} = -\sigma_{k'i}\cdot\vec{u_i} \tag{8.5}$$

$$\frac{\partial O_{ij}}{\partial \vec{u_i}} = (1-\sigma_{j'i})\cdot\vec{u_i} - \sum_{k=1}^{K}\sigma_{k'i}\cdot\vec{u_i} \tag{8.6}$$

この更新式において，8.4 式は文脈ベクトルの更新式を表し，8.5 はネガティブサンプリング中にサンプルされた文脈ベクトル対する更新式を表し，8.6 は頂点ベクトルの更新式を表す．このそれぞれの更新式を元に算出した値に，SGD 法の更新方法に従い，学習率 $\rho_t = \rho_0(1-\frac{t}{T})$ を乗じたものをそれぞれのベクトルに加算することで更新を行う．

8.2 接続モデルの更新式

本章で提案している，“移動目的推定モデル” は，地理的制約グラフと目的近接性のグラフが人流グラフを決定づけるというものである．つまり，ここでは，３つのネットワークをそれぞれ異なる空間へ写像することを目的としている．この３つのネットワークを最適化するための目的関数を，以下のように定める．

$$O = O_{sc} + O_{sr} + O_{ss} \tag{8.7}$$

それぞれのネットワークに設定された目的関数は，接続モデルの場合，相互に密接には依存していない．そこで，本章ではそれぞれの目的関数を順番に最適化する方法を採った．つまり，接続モデルの更新式は，地理的制約グラフ (G_{sc}) の目的関数 O_{sc}，目的近接性グラフ (G_{sr}) 上の目的関数 O_{sr}，人流グラフ (G_{ss}) 上の目的関数 O_{ss} を設定し，それぞれのグラフ上の頂点を表すベクトルとして更新される．そのため，それぞれのグラフ上での更新式は，8.4 式〜8.6 式と同様となる．ただし，接続モデルにおいては，G_{ss} 上で更新した分散表現は，G_{sc}, G_{sr} 上の分散表現を前半部分，後半部分として接続したものとしている．よって，G_{ss} 上で更新した部分はそれぞれのグラフ上の該当部分を更新する．

8.3 内分モデルの更新式

8.7 式で示したとおり，本章で提案したモデルは複数のネットワークを別々のベクトル空間へ写像するモデルである．しかし，内分モデルにおいては G_{ss} はそれ以外

のグラフと密接に依存している．この依存関係を表したものが 8.8 式である．

$$\vec{u}_{ss_i} = \alpha\vec{u}_{sc_i} + (1-\alpha)\vec{u}_{sr_i} \tag{8.8}$$

この内分モデルでは，G_{sc}, G_{sr} に関しては，接続モデルと同様に 8.4 式〜8.6 式で更新される．ただし，G_{ss} に関しては，この 8.8 式を用いて，それぞれのグラフ上の該当部分の更新も行う．この G_{ss} 上の更新式は，8.8 式を 8.1 式〜8.3 式に代入し，それぞれのベクトルに関して微分を行うことで導出される．

まず，文脈ベクトルに関しては以下のような更新式になる．

$$\frac{\partial O_{ij}}{\partial \vec{u_j'^{sc}}} = (1-\sigma_{j'i})\cdot\alpha\cdot\vec{u_i^{ss}} \tag{8.9}$$

$$\frac{\partial O_{ij}}{\partial \vec{u_j'^{sr}}} = (1-\sigma_{j'i})\cdot(1-\alpha)\cdot\vec{u_i^{ss}} \tag{8.10}$$

$$\tag{8.11}$$

ネガティブサンプリングによって，選択される頂点の文脈ベクトルに関しては，以下のような更新式となる．

$$\frac{\partial O_{ij}}{\partial \vec{u_k'^{sc}}} = \sigma_{k'i}\cdot\alpha\cdot\vec{u_i^{ss}} \tag{8.12}$$

$$\frac{\partial O_{ij}}{\partial \vec{u_k'^{sr}}} = \sigma_{k'i}\cdot(1-\alpha)\cdot\vec{u_i^{ss}} \tag{8.13}$$

$$\tag{8.14}$$

最後に，頂点ベクトルの更新式は以下のようになる．

$$\frac{\partial O_{ij}}{\partial \vec{u_i^{sc}}} = (1-\sigma_{j'i})\cdot\alpha\cdot\vec{u_i^{ss}} - \sum_{k=1}^{K}\sigma_{k'i}\cdot\alpha\cdot\vec{u_i^{ss}} \tag{8.15}$$

$$\frac{\partial O_{ij}}{\partial \vec{u_i^{sr}}} = (1-\sigma_{j'i})\cdot(1-\alpha)\cdot\vec{u_i^{ss}} - \sum_{k=1}^{K}\sigma_{k'i}\cdot(1-\alpha)\cdot\vec{u_i^{ss}} \tag{8.16}$$

最後に，それぞれのベクトル間の内分比の最適化を以下の更新式を用いて行う．

$$\begin{aligned}\frac{\partial O_{ij}}{\partial \alpha} = (1-\sigma_{j'i})\\ \{(\vec{u_j'^{sc}} - \vec{u_j'^{sr}})\cdot\vec{u_i^{ss}} - \vec{u_j'^{ss}}\cdot(\vec{u_i^{sc}} - \vec{u_i^{sr}})\}\\ -\sum_{k=1}^{K}\sigma_{k'i}\\ \{(\vec{u_k'^{sc}} - \vec{u_k'^{sr}})\cdot\vec{u_i^{ss}} - \vec{u_k'^{ss}}\cdot(\vec{u_i^{sc}} - \vec{u_i^{sr}})\}\end{aligned} \tag{8.17}$$

付録Ｂ:第6.6節で取得したアンケート

8.4　ワークショップ

第６章では，システムを開発し，関西の鉄道運営会社６社局の担当者の方に集まってもらい，システムを使ってもらうワークショップを開催した．ワークショップは，2016 年 09 月 05 日に開催し，この６社局に在籍する担当者 14 名が参加した．各社局ごとの参加者はそれぞれ１名～5 名で，各社局ごとに１台のクライアントＰＣを用意した．ワークショップは次の手順で進めた．まず，始めの３０分でワークショップの位置づけ，開発システムの機能，使用方法の説明を行った．次に各社局ごとに１つのチームになっていただきそれぞれのチームに１台ずつＰＣを提供した．そして１時間３０分ほどかけて，チームとして自由に分析を行う時間とした．そしてその後の３０分で，分析結果をチームごとに発表してもらった．分析内容は，各チームごとに自由に発想してもらい，議論を行い，最終的にパワーポイントの資料としてまとめ，発表してもらった．

今回開発したシステムでは，それぞれの社局が運営する路線内の駅に対して，自社局と他社局からの乗降客の移動状況を調べることができる．通常，各社局の担当者は，自社局が運営する路線内での乗客の移動のみ調べることができるが，他社局内の乗客の移動について調べることができない．第６章で開発したシステムは，他社局内の駅と乗降客を通じてどのような関係性があるのか見れる点で彼らにとって画期的で，強い関心を持って取り組んでもらえた．

ここで，各社がそれぞれに分析した内容について，簡単に紹介する．発表された分析結果は全部で 13 件であった．内訳は，駅周辺で開催されたお祭りなどのイベントによる移動量に関するものが４件，外国からの旅行者の移動に関するものが３件，割引キャンペーン，運賃改定，競合他社の駅設置の効果測定をしようというものが３件，駅周辺の商業施設による移動を調べようとするものが２件，他社局との乗り換え駅の状況に関するものが１件であった．これらの結果から多くの社局において，イベントによる需要喚起や外国人旅行客の動向に強い関心を持っていることがわかる．実際の分析結果の一部は本文中で説明したが，普段担当している方ならではの視点を持った分析が行われ，それぞれに興味深いもので，本システムを利用することで，これらの分析を定量的に行える点で有効である．

8.5　ワークショップ後に配布した質問票

ワークショップ終了後，システムに関する評価を得るために質問票に記入してもらった．この質問票は，ワークショップ参加者のうち，各社局の代表者の方１名に記入してもらったものである．実際の質問票を図 8.1 に示す．この質問票は，他社局の乗降データとの統合による価値，分析システムの有用性，機密漏洩に対するリスクの３つの項目に関するもので，各質問は５段階のリッカート尺度を用いて数値的に評価するものと自由に記述してもらうもので構成されている．第 6 章では，このアンケートの回答結果を集計したものを用いて評価を行っている．

複数鉄道運営会社横断システムWSアンケート

所属会社 _____

名前 _____

1 あなたは普段から業務でデータを利用した分析やレポート作成を行っていますか ○ はい ○ いいえ

2 今回利用しました6社局広域データの利用価値はどの程度と感じましたか

○ 5(非常に価値がある) ○ 4(充分に価値がある) ○ 3(価値がある)

○ 2(あまり価値が無い) ○ 1(まったく価値が無い)

3 （前問で，2,1を答えた方のみ）理由をお答えください（例）自社局のデータのみで充分，駅周辺の商業施設のPOSデータが必要，等

4 今回試作したシステム（駅乗降客分析システム）の有効性はどの程度と感じましたか

○ 5(非常に価値がある) ○ 4(充分に価値がある) ○ 3(価値がある)

○ 2(あまり価値が無い) ○ 1(まったく価値が無い)

5 （前問で，2,1を答えた方のみ）理由をお答えください（例）駅での乗降を分析するものでは利用シーンが無い，等

6 認証機能によって自社局内のみの駅，路線のみが検索対象になったことによって自社局の乗降データに関する機密は保護されていると感じますか ○ はい ○ いいえ

7 （「いいえ」の方のみ）懸念している機密情報とは何ですか

図 8.1: ワークショップ終了後のアンケートに利用した質問票.

参考文献

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[2] Mislove Alan, Lehmann Sune, Ahn Yong-Yeol, Onnela Jukka-Pekka, and J. Niels Rosenquist. *Understanding the Demographics of Twitter Users*, pages 554–557. AAAI Press, 2011.

[3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[4] Loulwah AlSumait, Daniel Barbar, and arlotta Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *2008 Eighth IEEE International Conference on Data Mining*, pages 3–12, Dec 2008.

[5] Kevin Ashton. That ‘internet of things’ thing. *RFiD Journal*, 22(7):97–114, 2009.

[6] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, 2011.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[8] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[9] Tom Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35, 2010.

[10] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[11] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 891–900, New York, NY, USA, 2015. ACM.

[12] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA, 2010. ACM.

[13] Aron Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA, 2010. ACM.

[14] Boyd Danah and Crawford Kate. Critical questions for big data. *Information, Communication & Society*, 15(5):662–679, 2012.

[15] Inderjit S. Dhillon and Suvrit Sra. Generalized nonnegative matrix approximations with bregman divergences. In *Neural Information Processing Systems (NIPS)*, dec 2005.

[16] Francis X Diebold. On the origin(s) and development of the term "big data". *PIER Working Paper*, 12(037), 2012.

[17] Bass M. Frank. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969.

[18] R. Edward Freeman. Strategic management: A stakeholder approach. *Advances in strategic management*, 1(1):31–60, 1983.

[19] Anindya Ghose and Panagiotis G. Ipeirotis. Designing novel review ranking systems: Predicting the usefulness and impact of reviews. In *Proceedings of the Ninth International Conference on Electronic Commerce*, ICEC '07, pages 303–310, 2007.

[20] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[21] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Rec.*, 42(2):17–28, July 2013.

[22] Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[23] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 174–181, 1997.

[24] Jing He, Xin Li, Lejian Liao, Dandan Song, and William K. Cheung. Inferring a personalized next point-of-interest recommendation model with latent behavior patterns. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 137–143. AAAI Press, 2016.

[25] Masahiko Itoh, Daisaku Yokoyama, Masashi Toyoda, Yoshimitsu Tomita, Satoshi Kawamura, and Masaru Kitsuregawa. Visual fusion of mega-city big data: An application to traffic and tweets data analysis of metro passengers. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 431–440, Oct 2014.

[26] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[27] Cannarella John and A. Spechler Joshua. Epidemiological modeling of online social network dynamics. *CoRR*, abs/1401.4208, 2014.

[28] Elihu Katz and F. Lazarsfeld Paul. Personal influence. *The Part Played by People in the Flow of Mass Communication. New York*, 1955.

[29] Kevin Kelly. *The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future*. Penguin Publishing Group, 2016.

[30] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 115, pages 700–721, 1927.

[31] Jan Kietzmann and Ana Canhoto. Bittersweet! understanding and managing electronic word of mouth. *Journal of Public Affairs*, 13(2):146–159, 2013.

[32] Dohyeong Kim, Malabika Sarker, and Priyanka Vyas. Role of spatial tools in public health policymaking of bangladesh: opportunities and challenges. *Journal of Health, Population and Nutrition*, 35(1):1–5, 2016.

[33] Philip Kotler and Ronald E. Turner. *Marketing management: analysis, planning, and control*. Prentice-Hall, Upper Saddle River, NJ, USA, 1979.

[34] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, New York, NY, USA, 2015. ACM.

[35] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010.

[36] Neal Lathia and Licia Capra. How smart is your smartcard?: Measuring travel behaviours, perceptions, and incentives. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 291–300, New York, NY, USA, 2011. ACM.

[37] Paul F. Lazarsfeld, Bernard Berelson, and Hazel Gaudet. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press, 1948.

[38] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. Google flu trends still appears sick: An evaluation of the 2013-2014 flu season. *Available at SSRN 2408560*, 2014.

[39] Maurizio Lenzerini. Data integration: A theoretical perspective. pages 233–246, 2002.

[40] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185. Curran Associates, Inc., 2014.

[41] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[42] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 194–200. AAAI Press, 2016.

[43] Yu Liu, Chaogui Kang, Song Gao, Yu Xiao, and Yuan Tian. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, 14(4):463–483, 2012.

[44] Ying Long and Zhenjiang Shen. *Geospatial Analysis to Support Urban Planning in Beijing*, chapter Discovering Functional Zones Using Bus Smart Card Data and Points of Interest in Beijing, pages 193–217. Springer International Publishing, Cham, 2015.

[45] Rong Lu and Qing Yang. Trend analysis of news topics on twitter. *International Journal of Machine Learning and Computing*, 2(3):327, 2012.

[46] Paolo Mariti and Robert H Smiley. Co-operative agreements and the organization of industry. *The Journal of industrial economics*, pages 437–451, 1983.

[47] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[48] Matt Mohebbi, Dan Vanderkam, Julia Kodysh, Rob Schonberger, Hyunyoung Choi, and Sanjiv Kumar. Google correlate whitepaper. *Web document: correlate. googlelabs. com/whitepaper. pdf. Last accessed date: August*, 1:2011, 2011.

[49] Seth A. Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 33–41, New York, NY, USA, 2012. ACM.

[50] Masanao Ochi, Yutaka Matsuo, Makoto Okabe, and Rikio Onai. Rating prediction by correcting user rating bias. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 452–456, Washington, DC, USA, 2012. IEEE Computer Society.

[51] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. 14:1532–1543, 2014.

[52] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710, New York, NY, USA, 2014. ACM.

[53] Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Proceedings of the First International Conference on Human Behavior Understanding*, HBU'10, pages 14–25, Berlin, Heidelberg, 2010. Springer-Verlag.

[54] Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59, 2013.

[55] Thomas S. Robertson. Consumer innovators: The key to new product success. *California Management Review*, 10(2):23–30, 1967.

[56] Camille Roth, Soong Moon Kang, Michael Batty, and Marc Barthélemy. Structure of urban movements: polycentric activity and entangled hierarchical flows. *PloS one*, 6(1):e15923, 2011.

[57] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

[58] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 285–295, New York, NY, USA, 2001. ACM.

[59] David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. Peaks and persistence. page 355, 2011.

[60] Lijun Sun and Jian Gang Jin. Modeling temporal flow assignment in metro networks using smart card data. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 836–841, Sept 2015.

[61] Toshimitsu Takahashi, Ryota Tomioka, and Kenji Yamanishi. Discovering emerging topics in social streams via link anomaly detection. pages 1230–1235, 2011.

[62] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. Rumor diffusion and convergence during the 3.11 earthquake: A twitter case study. *PLoS ONE*, 10(4):e0121443, 04 2015.

[63] Jian Tang, Meng Qu, and Qiaozhu Mei. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1165–1174, New York, NY, USA, 2015. ACM.

[64] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1067–1077, New York, NY, USA, 2015. ACM.

[65] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proc. of ACL*, pages 417–424, 2002.

[66] JD Ullman. Information integration using logical views. *Theoretical Computer Science*, 2000.

[67] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

[68] Yingzi Wang, Nicholas Jing Yuan, Defu Lian, Linli Xu, Xing Xie, Enhong Chen, and Yong Rui. Regularity and conformity: Location prediction using heterogeneous mobility data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1275–1284, New York, NY, USA, 2015. ACM.

[69] Zih Syuan Wang, Jing Fu Juang, and Wei Guang Teng. Predicting poi visits with a heterogeneous information network. In *2015 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 388–395, Nov 2015.

[70] Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, and Rikio Onai. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2541–2544, New York, NY, USA, 2011. ACM.

[71] Duncan J. Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton University Press, Princeton, NJ, USA, 1999.

[72] Xindong Wu, Xingquan Zhu, Gong Qing Wu, and Wei Ding. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):97–107, Jan 2014.

[73] Cheng Yang, Maosong Sun, Wayne Xin Zhao, Zhiyuan Liu, and Edward Y. Chang. A neural network approach to joint modeling social networks and mobile trajectories. *CoRR*, abs/1606.08154, 2016.

[74] Daisaku Yokoyama, Masahiko Itoh, Masashi Toyoda, Yoshimitsu Tomita, Satoshi Kawamura, and Masaru Kitsuregawa. A framework for large-scale train trip record analysis and its application to passengers' flow prediction after train accidents. In *Advances in Knowledge Discovery and Data Mining*, pages 533–544. Springer, 2014.

[75] Quan Yuan, Gao Cong, and Aixin Sun. Graph-based point-of-interest recommendation with geographical and temporal influences. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 659–668, New York, NY, USA, 2014. ACM.

[76] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. Cambridge University Press, New York, NY, USA, 2014.

[77] Fan Zhang, Juanjuan Zhao, Chen Tian, Chengzhong Xu, Xue Liu, and Lei Rao. Spatiotemporal segmentation of metro trips using smart card data. *IEEE Transactions on Vehicular Technology*, 65(3):1137–1149, March 2016.

[78] Fuzheng Zhang, Nicholas Jing Yuan, Yingzi Wang, and Xing Xie. Reconstructing individual mobility from smart card transactions: A collaborative space alignment approach. *Knowl. Inf. Syst.*, 44(2):299–323, August 2015.

[79] Shenglin Zhao, Tong Zhao, Irwin King, and Michael R. Lyu. GT-SEER: geo-temporal sequential embedding rank for point-of-interest recommendation. *CoRR*, abs/1606.05859, 2016.

[80] Yu Zhao, Zhiyuan Liu, and Maosong Sun. Representation learning for measuring entity relatedness with rich information. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1412–1418. AAAI Press, 2015.

[81] Yu Zheng. Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol.*, 6(3):29:1–29:41, May 2015.

[82] Yu Zheng, Yanchi Liu, Jing Yuan, and Xing Xie. Urban computing with taxicabs. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 89–98, New York, NY, USA, 2011. ACM.

[83] Feng Zhenni and Yanmin Zhu. A survey on trajectory data mining: Techniques and applications. *IEEE Access*, 4:2056–2067, 2016.

[84] Paul Zikopoulos, Chris Eaton, et al. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.

[85] 情報通信総合研究所 株式会社. ビッグデータの流通量の推計及びビッグデータの活用実態に関する調査研究. 2015.

[86] 遠藤 薫. 「ネット世論」という曖昧:〈世論〉,〈小公共圏〉,〈間メディア性〉(〈特集〉世論と世論調査). **マス・コミュニケーション研究**, (77):105–126, 2010.

[87] 遠藤 薫. 「序章　なぜいまジャーナリズムを考えるか. **間メディア社会の＜ジャーナリズム＞　ソーシャルメディアは公共性を変えるか**, pages 1–17, 2014.

[88] 経済産業省.「需要予測の精度向上による食品ロス削減及び省エネ物流プロジェクト」の最終報告. **官公庁環境専門資料**, 50(3):91–94, may 2015.

[89] 公正取引委員会競争政策研究センター, 穂高 森田, 秀弥 林, 弘毅 荒井, and 元 宏西村. **企業の提携・部分的結合に関する研究**. Number CR03-10 in 競争政策研究センター共同研究. Competition Policy Research Center, Japan Fair Trade Commission, 2010.

[90] 石井 晃 and 吉田 就彦. ヒット現象の数理モデル. **鳥取大学工学部研究報告**, 36(39):71–80, 2005.

業績

第2章に関連する業績

査読付き国際会議

- Rating Prediction by Correcting User Review Bias.
  Masanao Ochi, Yutaka Matsuo, Makoto Okabe, Rikio Onai.
  Proceedings of the 2012 IEEE/WIC/ACM Web Intelligence (WI 2012), IEEE Press, Macau, China, pp. 452-456, Dec. 2012.
- Rating Prediction using Feature Words Extracted from Customer Reviews.
  Masanao Ochi, Makoto Okabe, Rikio Onai.
  Proceedings of the 34th Annual International ACM SIGIR Conference (SIGIR 2011), pp.1205-1206, Jul. 2011.
- Jasmine: Real-time Local Event Detection System by Propagating Geolocation Information to Microblogs.
  Kazufumi Watanabe, Masanao Ochi, Makoto Okabe, Rikio Onai.
  Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM 2011), pp.2541-2544, Oct. 2011.

査読なし国内会議

- レビュー文を利用したランキング関数の特徴量の提案.
  大知 正直，岡部 誠，尾内 理紀夫.
  人工知能学会 第 4 回 情報編纂研究会 (JSAI-SIG-IC), Feb. 2011.

第4章に関連する業績

査読付き和文論文誌

- 口コミ指数による事例類型化に基づく複数メディアのヒット前の露出を先行指標とした情報拡散過程の分析.

大知 正直，長濱 憲，榊 剛史，森 純一郎，坂田 一郎.
広報研究，第 20 号，pp. 35-51, March. 2016.

受賞

- 2016 年 10 月 日本広報学会 研究奨励賞.
  大知 正直, 長濱 憲, 榊 剛史，森 純一郎，坂田 一郎.
  日本広報学会 第 22 回研究発表全国大会, 北海道大学大学院国際広報メディア・観光学院, 札幌, Oct. 2016.

第 5 章に関連する業績

査読付き英文論文誌

- Geospatial Area Embedding Based on the Movement Purpose Hypothesis using Large-scale Mobility Data from Smart Card.
  Masanao Ochi，Yuko Nakashio, Matthew Ruttley, Junichiro Mori, Ichiro Sakata.
  International Journal of Communications, Network and System Sciences (IJCNS), Vol.9, No.11, pp. 519-534. Nov. 2016.

査読付き国際会議

- Representation learning for geospatial areas using large-scale mobility data from smart cards.
  Masanao Ochi, Yuko Nakashio, Yuta Yamashita, Ichiro Sakata, Kimitake Asatani, Matthew Ruttley, Junichiro Mori, Ichiro Sakata.
  Papers from the 5th International Workshop on Pervasive Urban Applications in conjunction with ACM UbiComp 2016(PURBA2016), Heidelberg, Germany, pp. 1381-1389, Sep. 2016.

査読なし国内会議

- 空間的依存性を考慮したネットワークエンベディング手法の提案
  大知 正直，浅谷 公威，森 純一郎，坂田 一郎.
  人工知能学会 第 30 回 全国大会, Jun. 2016.

講演・シンポジウム 等

- 空間的依存性を考慮したネットワークエンベディング手法の提案.
  大知 正直.
  次年度の人工知能学会 近未来チャレンジセッションへの参加求む！ サバイバル 世界価値観データベースに基づく世界消費者の把握, 大阪市中央公会堂, 大阪, Oct. 2016.

第6章に関連する業績

査読付き和文論文誌

- マルチステークホルダーによる相互データ提供に基づいた意思決定支援システム.
  大知 正直，前田 ニコラス 高志, 浅谷 公威, 鳥海 不二夫, 森 純一郎, 坂田 一郎.
  (投稿準備中).

その他

査読付き国際会議

- Understanding Rating Behaviour and Predicting Ratings by Identifying Representative Users.
  Rahul Kamath, Masanao Ochi, Yutaka Matsuo.
  Papers from The 29th Pacific Asia Conference on Language, Information and Computation (PACLIC2015), Shanghai, China, Nov. 2015.
- Discovering Behavior Patterns from Social Data for Managing Personal Life.
  Rui Pan, Masanao Ochi, Yutaka Matsuo.
  Papers from the 2013 AAAI Spring Symposium, Palo Alto, California, USA, pp. 25-27, March. 2013.

査読なし国内会議

- 移動データからのグループダイナミクスの分析.
  浅谷 公威，森 純一郎，大知 正直，坂田 一郎.
  人工知能学会 第30回 全国大会, Jun. 2016.

- 外国人旅行客の日本滞在時の環境向上のための分析.
  大知 正直，森 純一郎，坂田 一郎.
  観光情報学会 第 12 回 全国大会, Jun. 2015.
- 日米スタートアップのキーワードによるクラスタリングを用いた事業トレンド予測
  今井 響，大知 正直，松尾 豊.
  人工知能学会 第 29 回 全国大会, May. 2015.
- 推薦そのものがユーザに与える影響を考慮した情報推薦.
  大知 正直，関 喜史，川上 登福，小野木 大二，野村 眞平，吉永 恵一，松尾 豊.
  人工知能学会 第 28 回 全国大会, May. 2014.
- Web ニュースの主成分を用いたアルゴリズムトレード.
  宮崎 邦洋，大知 正直，松尾 豊.
  人工知能学会 第 28 回 全国大会, May. 2014.
- クラウドファンディングにおけるプロジェクトの資金調達可能性の分析.
  宮崎 邦洋，大知 正直，米良 はるか，松尾 豊.
  人工知能学会 第 27 回 全国大会, Jun. 2013.
- ユーザの成長を促進する情報推薦.
  大知 正直，関 喜史，川上 登福，小野木 大二，野村 眞平，吉永 恵一，松尾 豊.
  人工知能学会 第 27 回 全国大会, Jun. 2013.
- ソーシャルメディア上で自己増殖する人工生命の構築.
  大知 正直，松尾 豊.
  人工知能学会 第 27 回 全国大会, Jun. 2013.

謝辞

長く困難な５年間でありました．特に始めの不毛な３年間の後，退学させられそうになっている僕を快く研究室のメンバーとして受け入れていただいた坂田一郎教授に心より敬意と感謝を捧げたいと思います．また，博士課程において明確なビジョンもなく，その場その場で面白いと思った事象に飛びつく僕を生暖かく見守ってくださった，森純一郎准教授に深く感謝いたします．坂田一郎教授，森純一郎准教授には，博士論文執筆にあたり非常に重要なアドバイスを繰り返しいただきました．再度，御礼申し上げます．

「A friend in need is a friend indeed.（困った時の友こそ真の友）」という格言があります．まさにそのとおりで，最も苦しい時期に共に研究してくれた長濱憲さんには特に感謝します．共著で投稿した論文が採択され，研究奨励賞という賞もいただいたことで少し恩返しができたのではないかと思っております．榊剛史研究員には，データ提供や公私に渡るさまざまな相談にのっていただき本当に感謝しております．元はといえば，氏の書いた論文に衝撃を受けて本専攻を志したことでもあり，氏と交流できたことは僕にとって大きな財産であると思います．特に氏の博士論文と東京大学の松田尚子助教の博士論文は，今回の博士論文執筆において非常に参考にさせていただきました．本当にありがとうございます．

同様に暖かく仲間として迎え入れてくれた坂田・森研究室のメンバーにも深く感謝しております．特に加入当時のメンバーである中塩優子さん，山下雄大さん，丸井淳己氏，伊藤諒さん，岡太朗さん，安田洋介さん，田中和哉さんは，当時完全に意気消沈している僕に優しく接してくれました．また，時には食事をともに出来たのも良い思い出となっております．ありがとうございました．さて，忘れられないのは電気電子工学専攻に在籍していた田口直樹さんです．田口さんとは僕が辛い時期によく秋葉原のメイドカフェ[2]で一緒にお茶をしてくれました．また，一時期は寝食をともにし，上野広小路にある，“サウナ＆カプセルホテル ダンディ[3]”で何連泊もし，一緒に相互情報推薦に関する著書を電子書籍[4]として出版できたのはとても良い思い出です．今後また一緒に何か仕事ができるのを楽しみにしていますし，折り

[2]JAM Akihabara: http://jam-akiba.com/

[3]サウナ＆カプセルホテル ダンディ: http://www.u-excellent.com/dandy/index.html

[4]女性はあなたのココを見ている～ビッグデータから見えるモテ法則～: https://www.amazon.co.jp/dp/B00SI6IT60

にふれて飲食をともにしたいと思います．余談になりますが，僕はこの “林家ペー” 氏がイメージキャラクターを勤めている “サウナ＆カプセルホテル ダンディ” に会社員として働いていた頃から通算すると１０年近く宿泊していることになります．吉野家とカプセルホテルを一切利用しないでも全く懐具合が気にならない人生を送れるようになりたいとずっと思っていましたが，今のところ果たせる見込みはありません．

さて，本論文の中核の 1 つとなっている制約付き半教師あり学習による複数ネットワーク分散表現学習に関するアイデアは，現 Gunosy[5] の吉田宏司くんが誘ってくれた熱海旅行中に一緒にカフェ[6] で読んだ論文が元になりました．おかげさまで，そこでのアイデアを形にし，査読を経た上で発表することができたのは無上の喜びであります．当時一緒に旅行した人々にも同じくお礼を申し上げます．この吉田くんには，先日も六本木の “Shangri-La’s secret[7]” というお店で「きのこ鍋」をごちそうになりました．いつもありがとうございます．今後もよろしくお願いします．

また，博士論文を仕上げるにあたり校閲をしてくれた磯沼大くんにも，お礼を申し上げたいと思います．この磯沼くんには根津近辺のバー[8] で人生観について講釈を述べたりいたしましたが，なかなかの人格者でこれから立派になるのが楽しみな若者であります．

さて，僕自身は一人で集中して研究ができるタイプではありません．それぞれの研究でさまざまな形で協力してくれた方々のおかげでこのような形でまとめることができました．浅谷公威研究員には，研究だけでなく，自宅パーティに招いていただいたりしました．また，ドイツに研究発表[9] に赴いた際にも積極的に引率してくださいました．ありがとうございます．また，東京大学の鳥海不二夫准教授，前田ニコラス高志さんもつらい時期にやった研究を聞いていただき本当に助かりました．ありがとうございました．今後もよろしくおねがいします．

また，本論文にある研究はどれも実社会で蓄積されたデータがなければ研究ができませんでした．特に経営上重要なデータを提供していただいた株式会社スルッと KANSAI 様，株式会社ホットリンク様にお礼を申し上げます．

そもそもこのような研究に興味を持つきっかけとなったのは，東芝情報システム[10] という会社に在籍し，半導体設計技術者として働いている中で HSPICE[11] 用の回路情報の記述を自動化する仕事がきっかけだったと思います．実際の製品動作のシミュレーションに膨大なコストを投入していること，ハードウェアをすべてシンボ

[5] https://gunosy.co.jp/
[6] CAFE RoCA: http://caferoca.jp/
[7] Shangri-La’s secret: https://tabelog.com/tokyo/A1307/A130701/13175214/
[8] 例えば「くるみ」: https://tabelog.com/tokyo/A1311/A131106/13142739/
[9] UbiComp2016: http://ubicomp.org/ubicomp2016/
[10] https://www.tjsys.co.jp/
[11] https://www.synopsys.com/JP2/Tools/Verification/AMSVerification/CircuitSimulation/HSPICE/Pages/default.asp

ルとして抽象化することの２つに当時興味を持ち，半導体設計以外の分野でも取り組んでいきたいと考えたのが，博士取得への途方もない道のりのスタートだったように思います．一方で，当時日本の半導体産業の取り組み，グループ全体での人材活用に疑問を感じたことも動機の１つでした．

このような体験から一切ものづくりと関係の無い世界で，さまざまな事象を抽象化することで価値を創造していきたいと感じ電気通信大学の尾内教授（当時），岡部助教（当時）[12] の研究室に加えさせていただきました．修士課程時代にお世話になったこの研究室では，一切経済合理性を感じない研究であふれており，メーカーの技術職では味わえない自由なアイデアを考えることができました．この頃，お世話になった両先生には大変感謝しております．また，その時に得た同窓の友人たち（といっても 10 歳以上年下の若者たちですが）は今でも毎年夏に高尾山にハイキング [13] に行ったりして仲良くしていただいています．一方で当時の研究を結局論文誌としてまとめられなかった点は非常に申し訳なく思っております．自分の力不足であり，不明を恥じております．今後の研究活動で恩返しできればと思います．

僕は学部に８年在籍していました．当時大阪大学基礎工学部にいらっしゃった張紀久夫教授（当時）に「国立大学を卒業したものは国家予算を使って得た知識を社会に還元しなければならない」と当時研究活動に消極的であった僕に個人的に説諭していただいたことを思い出します．卒業後は，修士，博士課程で得た知識をできる限り社会に還元できるよう努力したいと思います．こうしたことが考えられるようになったのも，学部時代に所属していた菅滋正教授 [14]（当時）に厳しく指導していただき，学部を卒業させていただけたおかげだと思います．

最後に，これまで自分の思う道を進み，結局まわり道ばかりをしているような自分に対し，温かく見守りそして辛抱強く支援してくださった両親に対して深い感謝の意を表して謝辞を終えたいと思います．

[12] 岡部誠氏: http://makotookabe.com/
[13] 高尾山ビアマウント: http://www.takaotozan.co.jp/beermnt/
[14] http://decima.mp.es.osaka-u.ac.jp/sugaken/index_old.html