人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-01

# 家庭内センシングを簡易に実現する「おうちモニタキット」の構築とその活用に向けた検討

# Ouchi Monitor Kit: A Sensor Platform for Residential Monitoring

服部 俊一[1*]　三浦 輝久[1]　堤 富士雄[1]
Shunichi Hattori[1], Teruhisa Miura[1], Fujio Tsutsumi[1]

[1](一財) 電力中央研究所
[1]Central Research Institute of Electric Power Industry

**Abstract:** This paper introduces a sensor platform for residential monitoring, which is named "Ouchi Monitor Kit (OMK)". Researches and commercial services regarding residential monitoring have recently been popular thanks to smart meter installation and economical sensors. However, the study of residential monitoring is still an unfamiliar domain for researchers who specialize in computer science because it requires intimate understanding about devices and wireless communication. OMK is therefore developed to easily realize residential monitoring as an integrated platform including various sensor types such as electricity demand, temperature and so on. The examples and characteristics of collected sensor data are also introduced.

## 1　はじめに

本稿では，家庭内の電力消費量や温湿度，二酸化炭素濃度などのセンサデータを簡易に計測・収集可能なセンサキットである「おうちモニタキット」について紹介する．

スマートメータと呼ばれる，通信機能を持ち電力の利用状況（電力消費データ）をリアルタイムに計測できる次世代電力計の設置が全国で進められている．加えて，室内温度や玄関ドアの開閉など，家庭内の環境や行動を計測するセンサの低価格化・省電力化が進んでおり，家庭におけるセンサデータの簡易な収集・活用が可能となる環境が整備されつつある．

これらのデータには家庭内の状態や活動など様々な情報が含まれており，省エネや高齢者の見守りなど多くの用途への活用が期待できる．その一方で，これらのセンサデータの収集には計測機器や無線通信に関する知見，収集したデータの前処理・分析など，ハードからソフトに跨がる幅広い知識が必要となり，計算機科学など分野外の研究者・開発者にとって不慣れな領域と言える．

そこで，スマートメータや家庭内に設置したセンサから得られるデータを簡易に収集・表示可能なキットとして「おうちモニタキット (OMK)」を開発した（図 1）．OMK ではスマートメータから得られる電力消費データに加えて，室内気温や湿度，ドアの開閉などを計測するセンサに対応しており，搭載する機能やセンサの種類を継続的に変更・改善していくことができるよう設計している．


図 1: おうちモニタキット (OMK)

OMK はセンサを活用した家庭向けサービスの検討や，センサデータ分析技術の研究に必要なデータの簡易な収集を目的として開発を進めている．本稿では，OMK の構成や対応センサの紹介と共に，OMK を用いて計測したセンサデータの特徴について考察する．

*連絡先：(一財) 電力中央研究所
〒 240-0196 神奈川県横須賀市長坂 2-6-1
E-mail: shattori@criepi.denken.or.jp

## 2　家庭内センシングに関する動向

家庭内にセンサを設置して人の行動や状態を推定する試みは，学術研究を中心に進められてきた．これらの手法を本稿では「家庭内センシング」と表記する．例として，室内各部屋への人感センサと家電の ON/OFF 情報に基づいて異変状態検出を行う手法 [1] や，照度センサと電力計を用いて生活パターン推定を行う手法 [2] などが提案されている．

欧州や米国ではスマートメータの設置が日本に先んじて進められていることから，電力消費データのみを用いて家庭内センシングを行う手法が広く研究されている．居住者が在宅しているか否かを推定する「在・不在判定 (occupancy detection) [3]」と呼ばれる手法や，主幹の電力消費データから家電個別の利用状況を推定する「用途分解 (disaggregation) [4]」などが例として挙げられる．国内においてもスマートメータの設置が開始されたことを受けて，いくつかの事例が報告されている [5, 6]．電力消費データのみを用いる手法は「非侵入型 (non-intrusive)」のモニタリングと呼ばれ [7]，家庭内への機器設置を必要とせずに家庭内センシングを実現できることから費用や心理的負担という点で優位性がある．しかし，推定には高度な分析手法が必要になることや確実な推定は困難であることなどから，目的や制約条件に応じて他のセンサデータを組み合わせた分析が効果的と考える．

前述したスマートメータの設置に加えて，近年では市販の家庭向けセンサにおいて低価格化と省電力化が進んでいることもあり，一般家庭において電力を含むセンサデータの簡易な収集・活用が可能となる環境が整備されつつある．そのため，2016 年 4 月の電力小売全面自由化以降，競争環境にある電力業界でも顧客満足度向上を目的とした商用サービスの提供が始まっている[1]．

その一方で，センサ活用における技術的課題も無視できない．家庭内センシングを行うためには計測機器や無線通信に関する幅広い知見が必要であり，計測データを取得するまでの障壁が高い．環境を構築できたとしても，電波や電源，機器の信頼性の問題から継続的かつ高精度な計測が行えない場合も多い．市販センサを組合わせて簡易にデータ計測が行えるようになれば，より広い分野においてセンサデータの活用が期待できるが，市販センサの多くはスマートフォンからの閲覧のみで生データの収集が行えなかったり，マルチベンダ・クロスデバイスでの連携が困難であるといった課題が指摘されている [8]．そのため，家庭内センシングを簡易に実現するためには，様々なセンサデータを目的に合わせて組み合わることができるオープンなプラットフォームが必要と考える．

[1] https://www.service.tepco.co.jp/s/Anshin_Tooku/

## 3　おうちモニタキットの開発

### 3.1　開発要件と構成

OMK では電力消費データの計測を軸として，利用するセンサや機能を継続的に変更・改善するため以下に示す要件を満たすよう開発を進めている．

- スマートメータに接続して電力消費データを計測できる
- 開発者層の厚いハードウェア，ソフトウェアプラットフォームを採用する
- 家庭内で簡易に利用可能とすべく，Wi-Fi が無い環境でも利用可能とする
- 機能拡張や接続に制限のあるクローズドなプラットフォームでなく，可能な限りオープンなものを採用する
- センサの設置と継続的な利用が容易で，屋内の広い場所で安定して利用可能なセンサを採用する
- センサとゲートウェイ間の通信は一般住宅で安定して行える規格を採用する
- 可能な限り安価な部品・ソフトウェアを採用する

以上の要件に基づき選定した，OMK のハードウェア構成の一例を図 2 に示す．OMK 本体には安価かつ開発者コミュニティが充実している Raspberry Pi 3 Model B および Raspberry Pi 用 7 インチ公式タッチディスプレイを採用した．また，スマートメータとの通信にはローム社の Wi-SUN 通信用 USB ドングルである WSR35A1-00 を OMK 本体に内蔵し，これを用いて消費電力量の計測を行っている．ネットワークへの接続にはエイビット社のデータ通信端末 AK-020 を用いて Soracom Air による通信を行うこととした．


図 2: OMK のハードウェア構成の一例

表 1: OMK が現在対応しているセンサの一覧

| 製品名 | メーカー (販売元) | 計測データ | 通信プロトコル | 概要 |
|---|---|---|---|---|
| Wi-SUN USBドングル<br>WSR35A1-00 | ROHM | 電力 (主幹) | Wi-SUN<br>(ECHONET Lite) | 主幹電力値 (W, 計測間隔10秒～)<br>および30分間の積算電力消費量 (Wh) |
| Bluetoothワットチェッカー<br>REX-BTWATTCH1 | ラトックシステム | 電力 (コンセント) | Bluetooth | コンセントの消費電力値<br>(W, 計測間隔1秒～) |
| 温度センサ<br>STM431J, CS-EO431J他 | ROHM, アーミン他 | 温度 | EnOcean | 0-40℃の温度を1～30分間隔で計測 |
| マグネットセンサ<br>STM429J, STM250J他 | ROHM, アーミン他 | ドア開閉他 | EnOcean | ドアなどの開閉を検出 |
| ロッカースイッチ ESM210R他 | ROHM, アーミン他 | スイッチ押下 | EnOcean | スイッチの押下を検出 |
| 人感センサ HM92-01WHC | サイミックス | 人感 | EnOcean | 人体などの動きを検出 |
| Netatmoウェザーステーション<br>NET-OT-000001 | Netatmo | 温度, 湿度, CO2濃度,<br>気圧, 騒音 | Wi-Fi | 5分間隔で計測, Netatmo社のクラウドに<br>アップロードされたデータをAPIで取得 |

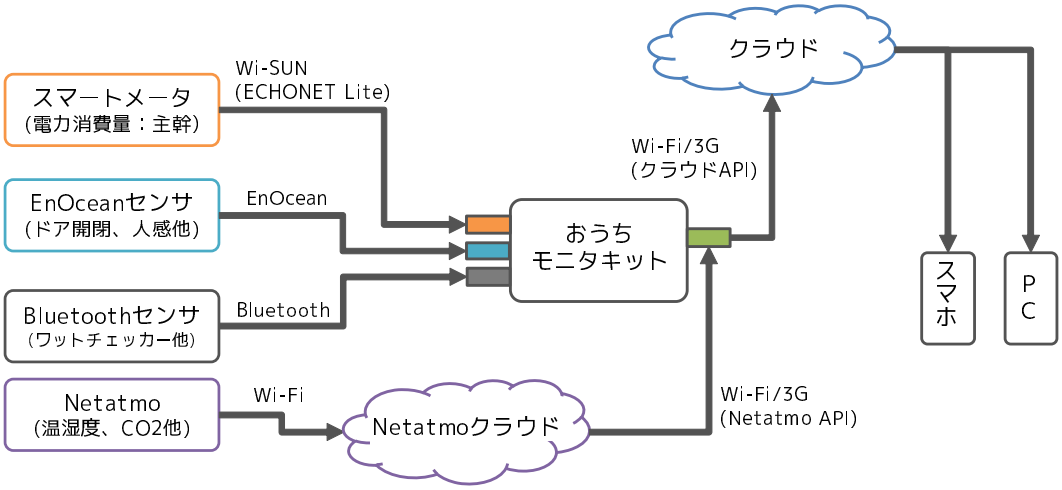



図 3: OMK における計測データの流れ

対応センサについては表 1 にまとめた．上記要件を満たす製品を調査した結果，EnOcean によるセンサを中心に採用することとした．EnOcean は電源がなくても動作するエナジーハーベスティング技術を用いており，920MHz 帯の安定した無線通信が行えること，設置環境の自由度が高いことが理由である．一方で，太陽光発電を用いたものがほとんどであることから暗所での動作に難があるなどいくつかの課題も明らかとなっており，今後は Bluetooth センサなど電池で動作する各種センサへの対応も進める予定である．

これらの機器は必要に応じて組み合わせることができる．例えば 3G でなく Wi-Fi を用いてネットワークに接続するのであれば，AK-020 および Soracom Air の利用は不要となる．利用するセンサも OMK の利用に必須なものはなく，必要なものだけ用意すればよい．

図 3 は，OMK における計測データの流れを図示したものである．計測されたセンサデータは OMK に集約され，その後必要に応じてクラウドにアップロードすることで PC などからセンサデータを閲覧することができる．なお，Netatmo ウェザーステーションのみ，製品仕様の都合から Netatmo 社のクラウド（製品利用者であれば無料で利用可）にアップロードされたものを API により取得している．

## 3.2 ディスプレイ表示

OMK では，計測されたセンサデータがリアルタイムに本体ディスプレイ上で表示される．OMK のディスプレイ表示例を図 4 に示す．表示項目や大きさは設定画面から自由にカスタマイズできる．各センサデータの値は数値と円の大きさ双方もしくは片方の形式により表示可能で，電力消費量のみ電気料金の目安を併記することができる．電気料金の算出は，表示されている消費量（瞬時値）が 1 時間続いた場合の金額とし，東京電力エナジーパートナー社が提供する一般的な料金プラン「従量電灯 B」における第 2 段階料金（1kWh=26.0 円）換算とした．ただし，電子レンジやドライヤーなど，短期間に瞬時値が跳ね上がる家電の場合は実際の使用量以上に高額の料金が表示されてしまうことから，電気料金の算出方法については改善の余地がある．

図 4 の例では，左側に現在の電力消費量および電気料金が大きく表示されており，右側に室内気温および

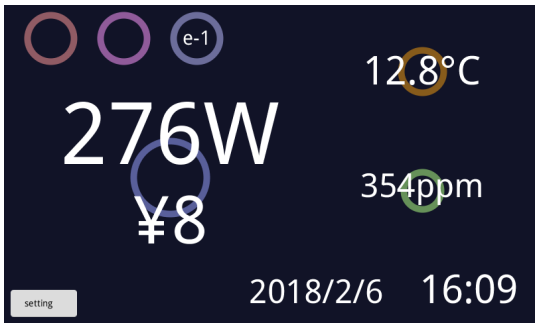



図 4: OMK のディスプレイ表示例

二酸化炭素濃度が表示されている．左上に並んでいる円は左からそれぞれマグネットセンサによるドア開閉，ロッカースイッチが押下されているか，人感センサに反応があったかという状態を表している．

# 4　おうちモニタキットによる計測例

本節では，OMK を家庭内に設置し，種々のセンサデータを計測した結果について，それぞれのセンサデータが持つ特徴と共に紹介する．

## 4.1　計測例 1

図 5 は，ある家庭において OMK を用いて計測したセンサデータ（電力，室内気温，人感，ドア開閉）をヒートマップ形式で可視化したものである．左側は電力消費量を，右側はその他のセンサデータを表示している．計測期間は 2016 年 7 月 21 日から 8 月 10 日までの 3 週間で，季節が夏季であることからエアコン利用により消費電力量と気温に強い関連が見られる．例えば 8 月 4 日から 6 日，8 日から 10 日にかけて 8 時〜12 時過ぎの時間帯は電力消費量が少なく，気温が徐々に上昇を続けている．ドア開閉・人感センサの反応も見られないことから，この時間帯は不在であると推定できる．また，13 時前後に帰宅し冷房を使用した結果として電力消費量が急上昇し，気温が下がったといった推測もできる．

在・不在判定への応用を考えた場合，電力消費データとドア開閉記録を組み合わせることでより高精度の推定が行えるだけでなく，より細かな時間解像度での外出/帰宅時刻の推定も可能と考えられる．この家庭ではマグネットセンサと人感センサを玄関付近に設置していることから，人感センサ反応後にドア開閉が確認されれば外出，その逆であれば帰宅と推定できる．しかし，今回の計測例では玄関付近の薄暗い場所に太陽光発電で動作する EnOcean センサを設置したことから，夜間・早朝を中心に欠測が発生することがあった．設置環境により適したセンサの選定や，他のセンサデータを組み合わせて欠測を補う手法の検討が必要と考える．

## 4.2　計測例 2

図 6 の例は別の家庭で計測したセンサデータ（電力，気温，二酸化炭素濃度，騒音）を折れ線グラフで表したものである．ここではワットチェッカーを用いてエアコンや冷蔵庫など家電単位での電力消費量も計測しているほか，気温・二酸化炭素濃度については室内の 2 箇所（リビング・寝室）で計測した．計測期間は 2017 年 7 月 22 日の 10 時から 24 時で，11 時半頃から 20 時過ぎまでは不在，それ以外は在宅となっている．

この家庭も計測時期が夏季であることから，外出時のエアコン停止により気温が上昇し，帰宅後エアコンの利用に伴って気温が下降していることがわかる．また，22 時過ぎにリビングから寝室へ移動し，それに伴って使用するエアコンも変更したことから，寝室へ移動後にリビングの気温が上昇し，寝室の気温が下降していることも観察できる．

二酸化炭素濃度は人の呼吸によって上昇することから，一般的には在宅時に上昇し不在時に下降する．図 6 では外出/帰宅，部屋の移動といった行動が濃度に反映されており，在・不在判定や居場所推定に有用と考えられる．一方で，追従が遅いこと，燃焼を伴う機器（ガスコンロ，石油ストーブ他）の利用や換気によって大きく値が変化するなど，いくつかの課題も存在する．

騒音は，一般的に人の活動や家電の利用により上昇する．図 6 の例においても，スピーカーや TV を利用する在宅時に上昇し，外出や寝室への移動後に下降していることがわかる．この結果から二酸化炭素濃度と同様，在・不在判定や居場所推定に有用と考えられる一方で，自動もしくはタイマー動作の家電や屋外の影響を受けるケースも想定する必要がある．今回の計測結果では外出時に床拭きロボットが壁に衝突することで騒音が発生したり，この家庭が線路に近いことから電車通過時に若干の騒音が発生した．

以上の結果から，それぞれのセンサデータには長所と短所，向き不向きが存在する．目的や制約条件に応じてセンサの最適な組合わせを考案する必要があり，そのためのプラットフォームとして OMK を活用できるのではないかと考える．

# 5　おわりに

本稿では家庭内センシングを簡易に実現可能なキットである OMK について紹介した．また，OMK を用いて計測されたセンサデータから，それぞれのデータが持つ特徴や用途について考察した．

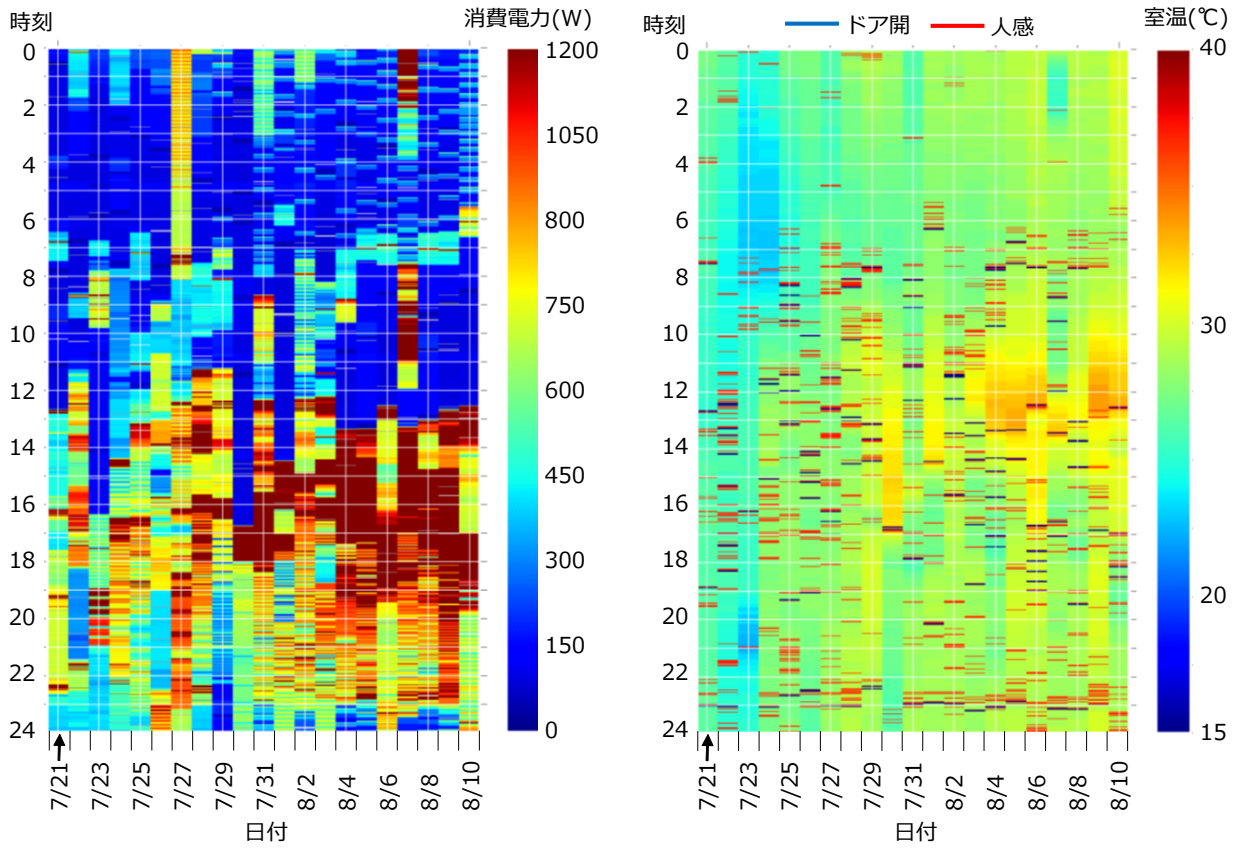



図 5: OMK による計測例 1 (電力，気温，人感，ドア開閉)

OMK はオープンなプラットフォームとして広く利用してもらうことを想定している．今後も対応センサやデータ処理に関する機能の追加を行うことで，より幅広い用途に活用できるよう改良を進めていく予定である．

## 参考文献

[1] 青木 茂樹, 大西 正輝, 小島 篤博, 福永 邦雄, 独居高齢者の行動パターンに注目した非日常状態の検出, 電気学会論文誌 E, Vol. 25, No. 6, pp. 259–265 (2005)

[2] S. Makonin and F. Popowich: “Home Occupancy Agent: Occupancy and Sleep Detection,” Journal on Computing, Vol. 2, No. 1, pp. 182–186 (2012)

[3] T. A. Nguyen and M. Aiello: “Energy intelligent buildings based on user activity: A survey,” Energy and Buildings, Vol. 56, pp. 244–257 (2013)

[4] K. C. Armel, A. Gupta, G. Shrimali and A. Albert: “Is disaggregation the holy grail of energy efficiency? The case of electricity,”Energy Policy, No. 52, pp. 213–234 (2013)

[5] 向井 登志広, 西尾 健一郎, 小松 秀徳, 内田 鉄平, 石田 恭子: スマートメータデータを活用した情報提供と行動変容－集合住宅におけるピーク抑制･省エネ実証事例－, 電力中央研究所研究報告, Y15002 (2016)

[6] 服部 俊一, 篠原 靖志: スマートメータデータからの実需要推定による在・不在判定の精度改善手法, 電気学会論文誌 C, Vol. 137, No. 9, pp. 1296–1303 (2017)

[7] G. W. Hart: “Nonintrusive appliance load monitoring,”Proceedings of the IEEE, Vol. 80, No. 12, pp. 1870–1891 (1992)

[8] 堤 富士雄, 三浦 輝久, 鶴見 剛也, 服部 俊一: 電気利用の拡大に向けた家庭内 IoT の動向と課題の整理, 電力中央研究所研究報告, R15012 (2016)

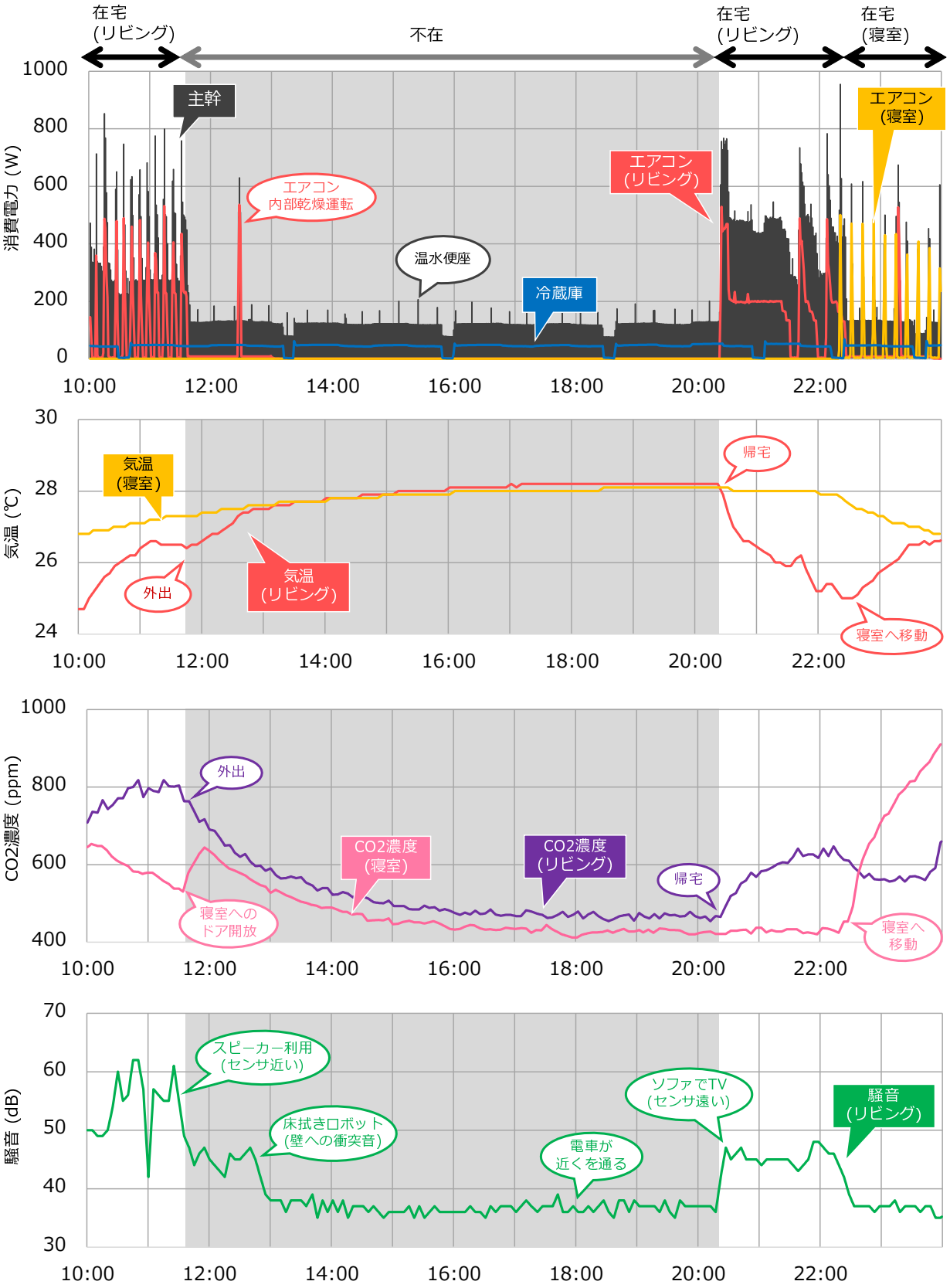



図 6: OMK による計測例 2 (電力，気温，二酸化炭素濃度，騒音)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-02

# 多次元軌跡データに対する類似部分軌跡検索の高速化

# Fast Similarity Search of Subtrajectories in Multidimensional Trajectory Databases

岡部 臨　　浦本 明　　尾崎 知伸 *
Nozomu Okabe　Akira Uramoto　Tomonobu Ozaki

日本大学 文理学部
College of Humanities of Sciences, Nihon University

**Abstract:** In the analysis of team sports, searching for simultaneous movements by a plurality of players specified by the user is one of basic operations. In this paper, we formulate this operation as a similarity search problem of subtrajectories in multidimensional trajectory data, and develop fast algorithms to solve the problem. The proposed algorithms are evaluated from the viewpoint of computation time and quality of obtained subtrajectories using real trajectory datasets on nine matches in Japanese professional football league.

## 1　はじめに

チームスポーツの分析において，利用者がクエリとして与えるフィールド上での動き（移動軌跡）と類似する場面を検索すること，すなわち移動軌跡の類似検索は基本的な操作の一つであり，近年盛んに研究が行われている．例えば Sha らは，バスケットボールにおけるパス・ショットなどの意味を含めた部分移動軌跡の検索を対象に，ハッシュ技術を用いた高速検索手法を提案している [1]．また Ohashi らは，動的属性を考慮した移動軌跡の類似検索手法を提案している [2]．動的属性とは，移動速度や周囲との位置関係など，時間経過に合わせて変化する（位置座標そのもの以外の）属性を表す．動的属性を考慮することで，より柔軟な類似検索を実現している．

既存研究の多くは，一つの移動軌跡をクエリとすることを前提としているが，チームスポーツにおける類似検索を考えた場合，複数人の移動軌跡から一人の移動軌跡を検索するだけでは，必ずしも十分であるとは言えない．なぜなら，チームスポーツにおいては，選手個人個人の動きだけではなく，しばしば複数人の選手が絡む連携プレーが重要となるからである．この問題を解決し，複数人の連携する動きを検索するには，クエリとして複数人の移動軌跡を与えると同時に，クエリに合致するプレイヤの組み合わせを考慮することが必要となる．

本論文ではこの問題を，多次元移動軌跡データに対する類似部分移動軌跡検索問題，すなわち長大な $N$ 次元移動軌跡データからクエリである $M$ 次元の移動軌跡と類似する部分を抽出する問題として定式化し，その高速アルゴリズムを提案する．また，サッカーにおける選手の移動軌跡データを対象に，計算時間と精度の面から提案手法を評価する．

*連絡先：日本大学文理学部情報科学科
〒 156-8550 東京都世田谷区桜上水 3-25-40
E-mail: tozaki@chs.nihon-u.ac.jp

## 2　準備

本章では準備として，移動軌跡に関する記法や定義を導入し，本論文で扱う問題を形式的に示す．

オブジェクト $o$ に関する移動軌跡 $tr^o$ を，時刻 $t$ における $o$ の位置座標 $(x_t^o, y_t^o)$ の系列

$$tr^o = (x_1^o, y_1^o), (x_2^o, y_2^o), \cdots, (x_t^o, y_t^o), \cdots, (x_T^o, y_T^o)$$

と定義する．また系列長 $T$ を $tr^o$ の長さと呼ぶ．なお本論文では，時刻 $t$ は 1 から始まり $T$ まで連続していることを仮定する．移動軌跡データ $tr^o$ における時刻 $i$ から $j\,(1 \leq i \leq j \leq T)$ の連続する部分移動軌跡を

$$tr_{i,j}^o = (x_i^o, y_i^o), \cdots, (x_j^o, y_j^o)$$

と表記する．

長さ $L$ の移動軌跡を要素とするサイズ $M$ の配列

$$Q = [tr^{q_1}, tr^{q_2}, \cdots, tr^{q_M}]$$

をクエリと呼ぶ．また配列 $Q$ の第 $i$ 要素を $Q[i]$ と表記する．一方，$N\,(N \geq M)$ 個のオブジェクト $d_1, d_2, \cdots, d_N$

に関する長さ $T\,(T >> L)$ の移動軌跡データの集合

$$D = \left\{ \begin{array}{rcl} tr^{d_1} & = & (x_1^{d_1}, y_1^{d_1}), \cdots, (x_T^{d_1}, y_T^{d_1}), \\ & \vdots & \\ tr^{d_N} & = & (x_1^{d_N}, y_1^{d_N}), \cdots, (x_T^{d_N}, y_T^{d_N}) \end{array} \right\}$$

を移動軌跡データベースと呼ぶ．移動軌跡データベース $D$ において，時刻 $i$ から長さ $L$，すなわち時刻 $i$ から $i+L-1$ までの部分移動軌跡の集合を

$$D_{(i,L)} = \left\{ tr_{i,i+L-1}^{d} \mid tr^d \in TR \right\}$$

と表記する．また，$d_1 \sim d_N$ から $M$ 個の異なる要素を選択し並べることで得られる配列を

$$\pi = [\, d_{x_1}, d_{x_2}, \cdots, d_{x_M} \,]$$

とし，$\pi$ の各要素 $\pi[j]$ に関する $D_{(i,L)}$ 中の部分移動軌跡を並べた配列を

$$D_{(i,L)}^{\pi} = [\, tr_{i,i+L-1}^{\pi[1]}, \cdots, tr_{i,i+L-1}^{\pi[M]} \,]$$

と表記する．

長さ $L$ の部分移動軌跡の配列である $D_{(i,L)}^{\pi}$ に対し，$x$ 軸 $y$ 軸それぞれの平均値と標準偏差

$$\begin{array}{l} \overline{X} = \frac{1}{L \times M} \sum_{i \le t \le i+L-1, 1 \le k \le M} x_t^{\pi[k]} \\ \overline{Y} = \frac{1}{L \times M} \sum_{i \le t \le i+L-1, 1 \le k \le M} y_t^{\pi[k]} \\ \sigma(X) = \sqrt{\frac{1}{L \times M} \sum_{i \le t \le i+L-1, 1 \le k \le M} (x_t^{\pi[k]} - \overline{X})^2} \\ \sigma(Y) = \sqrt{\frac{1}{L \times M} \sum_{i \le t \le i+L-1, 1 \le k \le M} (y_t^{\pi[k]} - \overline{Y})^2} \end{array}$$

を用いて，各座標位置を平均 0，分散 1 に標準化することで得られる正規化済み移動軌跡の配列を

$$n\_D_{(i,L)}^{\pi} = [\, n\_tr_{i,i+L-1}^{\pi[1]}, \cdots, n\_tr_{i,i+L-1}^{\pi[M]} \,]$$

where

$$\begin{array}{l} n\_tr_{i,i+L-1}^{d} = \left( \left( x_i^d - \overline{X} \right) / \sigma(X), \left( y_i^d - \overline{Y} \right) / \sigma(Y) \right), \\ \cdots, \left( \left( x_{i+L-1}^d - \overline{X} \right) / \sigma(X), \left( y_{i+L-1}^d - \overline{Y} \right) / \sigma(Y) \right) \end{array}$$

と表記する．同様に，移動軌跡の配列であるクエリ $Q$ に対する正規化済み移動軌跡の配列を $n\_Q$ と表記する．

正規化済みの移動軌跡データ配列 $n\_D_{(i,L)}^{\pi}$ とクエリ配列 $n\_Q$ との非類似度（距離）を，

$$dist(\, n\_D_{(i,L)}^{\pi},\, n\_Q \,) = \sum_{j=1}^{M} dtw(\, n\_D_{(i,L)}^{\pi}[j],\, n\_Q[j] \,)$$

と定義する．ここで $dtw$ は動的時間伸縮法 [3] に基づく距離（DTW 距離）であり，2 つの移動軌跡

$$\begin{array}{l} tr^d = (x_1^d, y_1^d), \cdots, (x_{L_d}^d, y_{L_d}^d) \\ tr^q = (x_1^q, y_1^q), \cdots, (x_{L_q}^q, y_{L_q}^q) \end{array}$$

に対し，

$$dtw(\, tr^d,\, tr^q \,) = \gamma(\, L_d,\, L_q \,)$$

where

$$\gamma(i,j) = \left\{ \begin{array}{ll} 0 & i = 0, j = 0 \\ \infty & i \neq 0, j = 0 \\ \infty & i = 0, j \neq 0 \\ d(i,j) + \min \begin{pmatrix} \gamma(i-1,j) \\ \gamma(i-1,j-1) \\ \gamma(i,j-1) \end{pmatrix} & \text{otherwise} \end{array} \right.$$

$$d(i,j) = \sqrt{(x_i^d - x_j^q)^2 + (y_i^d - y_j^q)^2}$$

と定義される．

本研究では，移動軌跡データベースに対する類似度上位 $K$ 検索を考える．すなわち，長さ $L$ の移動軌跡を保持する要素数 $M$ のクエリ配列 $Q$ と，$N\,(N \geq M)$ 個の各オブジェクトに関する長さ $T\,(T >> L)$ の移動軌跡の集合 $D$ に対し，$dist(n\_D_{(i,L)}^{\pi}, n\_Q)$ の値が上位 $K$ 番目以内である部分移動軌跡 $D_{(i,L)}^{\pi}$ を求める問題を対象とする．より形式的には，$K\,(K > 0)$ をパラメタとし，条件

$$\left| \{ (i', \pi') \mid dist(n\_D_{(i,L)}^{\pi}, n\_Q) > dist(n\_D_{(i',L)}^{\pi'}, n\_Q) \} \right| < K$$

を満たす $D_{(i,L)}^{\pi}$ を獲得する．

原理的には，移動軌跡データベースに対し，時刻 1 から $T-L+1$ までの各時刻から長さ $L$ の部分移動軌跡を取り出し，さらに $N$ 個あるオブジェクトから異なる $M$ 個を選択した上で，実際に動的時間伸縮法を用いて距離を計算する，という操作により，この問題の解を得ることが可能である．しかしこの場合，一回の動的時間伸縮法の計算量は 2 つの系列の長さの積となるので，全体として計算量は

$$\mathcal{O}\left( (T - L + 1) \times {}_N P_M \times L^2 \right)$$

となり，計算コストが非常に大きいことが問題となる．

## 3 最類似部分時系列の検索

これまでに，1 次元の時系列データを対象としたクエリに対する高速な類似検索手法として，動的時間伸縮法に基づくストリーム検索アルゴリズム The UCR suite[4] [1] が提案されている．本研究では，このアルゴリズムを出発点とし，多次元移動軌跡データへの拡張を行う．

[1] http://www.cs.ucr.edu/~eamonn/UCRsuite.html

SimilaritySearch( $D$, $Q$ )

$\langle bsf, loc\rangle := \langle\infty, -1\rangle$
$n\_Q :=$ normalize($Q$)
**for** $i$ in $\{1, \cdots, T-L+1\}$
  $n\_D_{(i,L)} :=$ normalize($D_{(i,L)}$)
  **if** $lb_{kim}LF(n\_D_{(i,L)}, n\_Q) > bsf \vee$
    $lb_{keogh}EQ(n\_D_{(i,L)}, n\_Q) > bsf \vee$
    $lb_{keogh}EC(n\_D_{(i,L)}, n\_Q) > bsf$
    **then continue**
  $dist := dtw(n\_D_{(i,L)}, n\_Q)$
  **if** $dist < bst$ **then** $\langle bsf, loc\rangle := \langle dist, i\rangle$
**end for**
**return** $D_{(loc,L)}$

図 1: The UCR suite の疑似コード

The UCR suite は，ストリームアルゴリズムの 1 種であり，大きく (1) 打ち切りを伴う標準化，(2) 下界計算による枝刈り，(3) 打ち切りを伴う DTW 計算から構成される．

検索対象である長さ $T$ の時系列データ $D$ とクエリである長さ $L$ の時系列データ $Q$ を対象とした，上位 1 位検索に対する The UCR suite の疑似コードを図 1 に示す．疑似コード中において，normalize は正規化を表す．また $lb_{Kim}LF$ は，$\mathcal{O}(1)$ で計算可能な DTW 距離の下界 [5]，$lb_{Keogh}EQ$ と $lb_{Keogh}EC$ は $\mathcal{O}(L)$ で計算可能な DTW 距離の下界 [6, 7] をそれぞれ表す．疑似コードが示す通り，開始位置を 1 から $T-L+1$ まで変化させながら，そのそれぞれでデータ $D_{(i,L)}$ の正規化と 3 種の下界計算及び DTW による距離計算を行い，最も類似する部分系列を特定する．このとき，複数の下界を計算量の小さい順に計算することで，少ない計算時間での枝刈り（DTW 計算の回避）を実現している．なお本疑似コードは概念レベルでの動作を表すものであり，実際のアルゴリズムではストリーム化及び下界計算を考慮したデータの部分的な正規化など，更なる最適化手法が利用されている．より詳細なアルゴリズムは原著論文 [4] を参照されたい．

# 4 類似部分移動軌跡の検索

本論文で対象とする，長さ $T$ の $N$ 次元移動軌跡データ $D$ に対する長さ $L$ の $M$ 次元クエリ $Q$ の類似部分移動軌跡検索問題では，距離 $dist(\, n\_D^{\pi},\, n\_Q\,)$ の計算が必要とされる [2]．距離 $dist(\, n\_D^{\pi},\, n\_Q\,)$ が DTW 距離の合計であることに着目し，The UCR suite に対して以下の 3 点の拡張を行うことで，多次元移動軌跡データに対する類似部分移動軌跡検索アルゴリズムを開発する．

1. 各 DTW 距離 $dtw(\, n\_D^{\pi}[j],\, n\_Q[j]\,)$ とその下界の計算を（1 次元の）時系列から（2 次元の）移動軌跡へと変更する
2. DTW 距離の合計である $dist(\, n\_D^{\pi},\, n\_Q\,)$ に対する下界を導入する
3. $N$ 次元の移動軌跡データ $D$ から，異なる $M$ 個を選択する順列 $\pi$ をすべて考慮する

以下では，上記の拡張に基づく基本アルゴリズム及びヒューリスティクスに基づく高速化アルゴリズムを導入する．

## 4.1 基本アルゴリズム

本論文における The UCR suite の拡張の一つである，時系列データから移動軌跡データへの変更に関しては自明な点も多く，また下界 $lb_{Keogh}$ に関してはその多次元化が既に報告されている [8]．従って，DTW 距離及びその下界計算に関しては，時系列データと移動軌跡データとで本質的な変更の必要はなく，本論文でも特に表記を分けることは行わない．

一方，$dist(\, n\_D^{\pi},\, n\_Q\,)$ が各要素における DTW 距離の合計であることと，不等式

$$0 \le lb_{Kim}LF \le \max(\, lb_{Keogh}EQ, LB_{Keogh}EC\,) \le dtw$$

が成り立つことより，$LB$ を DTW 距離の下界とし，また $S = \{1, 2, \cdots, M\}$，$S \supseteq S' \supseteq S''$ としたとき，

$$dist(S', n\_D^{\pi}, n\_Q) = \sum_{j \in S'} dtw(\, n\_D^{\pi}[j],\, n\_Q[j]\,)$$

$$dist(S', S'', n\_D^{\pi}, n\_Q) = \sum_{j \in S''} dtw(\, n\_D^{\pi}[j],\, n\_Q[j]\,) + \sum_{j \in S' \setminus S''} LB(\, n\_D^{\pi}[j],\, n\_Q[j]\,)$$

に対して，不等式

$$\begin{aligned} dist(n\_D^{\pi}, n\_Q) &\ge dist(S', n\_D^{\pi}, n\_Q) \\ &\ge dist(S', S'', n\_D^{\pi}, n\_Q) \end{aligned}$$

が成り立つ．この不等式より，$dist(S', n\_D^{\pi}, n\_Q)$ と $dist(S', S'', n\_D^{\pi}, n\_Q)$ は，$dist(n\_D^{\pi}, n\_Q)$ に対する下界を与えることとなるので，両者を $dist(n\_D^{\pi}, n\_Q)$ 計算に対する枝刈り基準として利用する．

この枝刈り基準を利用し，クエリの各要素に対して下界の計算と DTW 距離の計算を繰り返す上位 1 位類似部分移動軌跡検索のための基本アルゴリズムを図 2

[2] 簡略化のため以降ではデータ $n\_D^{\pi}$ に関する添字 $(i, L)$ を省略する．

Algorithm1( $D$, $Q$ )

1: $\langle bsf, loc, S\rangle := \langle \infty, -1, \emptyset\rangle$
2: $n\_Q :=$ normalize$(Q)$
3: **for** $i$ in $\{1, \cdots, T-L+1\}$
4: **for** $\pi$ in $Perm(D, Q)$
5: $n\_D^{\pi}_{(i,L)} :=$ normalize$(D^{\pi}_{(i,L)})$, $dist := 0$
6: **for** $j$ in $\{1, \cdots, M\}$
7: **if** $dist + lb_{kim}LF(n\_D^{\pi}_{(i,L)}[j], n\_Q[j]) > bsf \vee$
8: $dist + lb_{keogh}EQ(n\_D^{\pi}_{(i,L)}[j], n\_Q[j]) > bsf \vee$
9: $dist + lb_{keogh}EC(n\_D^{\pi}_{(i,L)}[j], n\_Q[j]) > bsf$
10: $dist + dtw(n\_D^{\pi}_{(i,L)}[j], n\_Q[j]) > bsf$
11: **then** $dist := \infty$, **break**
12: $dist := dist + dtw(n\_D^{\pi}_{(i,L)}[j], n\_Q[j])$
13: **end for**
14: **if** $dist < bst$ **then** $\langle bsf, loc, S\rangle := \langle dist, i, \pi\rangle$
15: **end for**
16: **end for**
17: **return** $D^{S}_{(loc,L)}$

図 2: 類似部分移動軌跡検索の基本アルゴリズム 1

に示す．なお図中において $Perm(D, Q)$ は，クエリ $Q$ の要素数である $M$ 個の異なる要素をデータ $D$ から選択する順列の全体集合である．また，アルゴリズム中で保持する解を上位 $K$ 件とすることで，同様の方法により上位 $K$ 位検索を実現することが可能である．

基本アルゴリズム（図 2）では，クエリの要素毎に下界の計算と DTW 距離の計算を繰り返す．しかし，計算の順序を変え，計算量の小さい下界の計算を優先的に行うことで，全体の計算時間を短縮することが期待できる．この考えに基づき，クエリの全要素に対して下界の計算を行った後に DTW 計算を行うアルゴリズムを図 3 に示す．

## 4.2 ヒューリスティクスの導入

本節では，チームスポーツへの応用の観点から，高速化のための 2 つのヒューリスティクスを導入する．

一つ目のヒューリスティクスは「似たようなプレイはしばしば似たような場所で起こる」という考えに基づき，考慮する順列 $\pi$ の順序（図 2, 図 3 の 4 行目）を制御するというものである．具体的には,（正規化前の）データ $D^{\pi}_{(i,L)}$ とクエリとの重心を計算し，その距離の昇順にデータの組み合わせ $\pi$ を考慮する．なお正規化済みデータ $n\_D^{\pi}_{(i,L)}$ を求めるためには $D^{\pi}_{(i,L)}$ の重心計算が必要となるため，このヒューリスティクスを利用する場合に係る追加コストは，順列集合 $P(D, Q)$ 内の要素を並べ替えるコストだけである．

Algorithm2( $D$, $Q$ )

1: $\langle bsf, loc, S\rangle := \langle \infty, -1, \emptyset\rangle$
2: $n\_Q :=$ normalize$(Q)$
3: **for** $i$ in $\{1, \cdots, T-L+1\}$
4: **for** $\pi$ in $Perm(D, Q)$
5: $n\_D^{\pi}_{(i,L)} :=$ normalize$(D^{\pi}_{(i,L)})$, $dist := 0$
6: **for** $j$ in $\{1, \cdots, M\}$
7: $dist := dist + lb_{kim}LF(n\_D^{\pi}_{(i,L)}[j], n\_Q[j])$
8: **if** $dist > bsf$ **then** $dist := \infty$, **goto** 22
9: **end for**
10: **for** $j$ in $\{1, \cdots, M\}$
11: $dist := dist - lb_{kim}LF(n\_D^{\pi}_{(i,L)}[j], n\_Q[j])$
$+lb_{keogh}EQ(n\_D^{\pi}_{(i,L)}[j], n\_Q[j])$
12: **if** $dist > bsf$ **then** $dist := \infty$, **goto** 22
13: **end for**
14: **for** $j$ in $\{1, \cdots, M\}$
15: $dist := dist - lb_{keogh}EQ(n\_D^{\pi}_{(i,L)}[j], n\_Q[j])$
$+lb_{keogh}EC(n\_D^{\pi}_{(i,L)}[j], n\_Q[j])$
16: **if** $dist > bsf$ **then** $dist := \infty$, **goto** 22
17: **end for**
18: **for** $j$ in $\{1, \cdots, M\}$
19: $dist := dist - lb_{keogh}EC(n\_D^{\pi}_{(i,L)}[j], n\_Q[j])$
$+dtw(n\_D^{\pi}_{(i,L)}[j], n\_Q[j])$
20: **if** $dist > bsf$ **then** $dist := \infty$, **goto** 22
21: **end for**
22: **if** $dist < bst$ **then** $\langle bsf, loc, S\rangle := \langle dist, i, \pi\rangle$
23: **end for**
24: **end for**
25: **return** $D^{S}_{(loc,L)}$

図 3: 類似部分移動軌跡検索の基本アルゴリズム 2

二つ目のヒューリスティクスは「大きさがクエリと極端に異なる検索結果は，実用上有益ではない」との考えに基づき，クエリと大きさが大幅に異なるデータ $D^{\pi}_{(i,L)}$ を検索対象から外すというものである．具体的には，クエリ $Q$ に対する幅 $W_Q$ と高さ $H_Q$ 及び $D^{\pi}_{(i,L)}$ に対する幅 $W_D$ と高さ $H_D$，すなわち

$$
\begin{aligned}
W_Q &= \max_{1\le k\le L, 1\le j\le M}(x_k^{q_j}) - \min_{1\le k\le L, 1\le j\le M}(x_k^{q_j}) \\
H_Q &= \max_{1\le k\le L, 1\le j\le M}(y_k^{q_j}) - \min_{1\le k\le L, 1\le j\le M}(y_k^{q_j}) \\
W_D &= \max_{i\le k\le i+L-1, 1\le j\le M}(x_k^{\pi[j]}) - \min_{i\le k\le i+L-1, 1\le j\le M}(x_k^{\pi[j]}) \\
W_Q &= \max_{i\le k\le i+L-1, 1\le j\le M}(y_k^{\pi[j]}) - \min_{i\le k\le i+L-1, 1\le j\le M}(y_k^{\pi[j]})
\end{aligned}
$$

に対し，$\alpha\,(0 \le \alpha \le 1)$ をパラメタとし，条件

$$
\begin{aligned}
(1-\alpha)W_Q \le W_D \le (1+\alpha)W_Q, \\
(1-\alpha)H_Q \le H_D \le (1+\alpha)H_Q
\end{aligned}
$$

を満たさない $D^{\pi}_{(i,L)}$ を計算対象から除外する．

表 1: 上位 10 位検索の計算時間（秒）

| $Q_{M,L}$ | $A_1$ | $A_2$ | $A_2^G$ | $A_2^R$ | $A_2^{GR}$ |
|---|---|---|---|---|---|
| $Q_{2,2}$ | 746 | 215 | 214 | 127 | 94 |
| $Q_{2,3}$ | 5994 | 1011 | 1014 | 674 | 237 |
| $Q_{3,2}$ | – | 1420 | 1753 | 1894 | 1821 |
| $Q_{3,3}$ | – | 3584 | 3963 | 3974 | 2627 |

表 2: 検索精度（$K$-適合率のマイクロ平均）

| $K$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| 10 | 0/100 | 2/100 | 0/100 | 1/100 |
| 20 | 0/200 | 2/200 | 0/200 | 1/100 |
| 30 | 1/300 | 2/300 | 0/300 | 1/100 |
| 40 | 3/400 | 2/400 | 0/400 | 2/400 |
| 50 | 3/500 | 2/500 | 2/500 | 2/500 |

# 5 評価実験

提案手法を評価するため，Java 言語を用いて各アルゴリズムを実装し，Windows PC（OS:Windows7 Pro, CPU:Intel Core-i3 2.40GHz, 主記憶:4GB）を用いて評価実験を行った．また実験には，J リーグのリーグ戦に関するデータ [3]（移動軌跡のサンプリングレートは 1 秒間 25 回）を利用した．

## 5.1 実行速度の評価

次元数（人数）$M$ 人，長さ $L$ 秒のクエリ $Q_{M,L}$ を用いて 5 試合を対象にそれぞれ上位 10 位検索を行い，その検索時間の平均値を計測した．実行結果を 表 1 に示す．表中において，$A_1$，$A_2$ はそれぞれ基本アルゴリズム 1（図 2）と 2（図 3）を表す．また，$A_2$ の上付き文字 $G$ と $R$ は，それぞれ重心 (center of gravity) によるヒューリスティクスと範囲 (range) によるヒューリスティクス（パラメタ $\alpha = 0.3$）を適用したことを表す．なお，“–” はタイムアウト（4 時間）を表す．

実験結果より，$A_1$ と比較し $A_2$ の方が高速であることが分かる．これは，計算量の少ない下界計算を優先させた効果によるものであると考えられる．また，$A_2$ と重心に基づく高速化を加えた $A_2^G$ を比較すると，$Q_{2,x}$ ではほぼ同じ，$Q_{3,x}$ では $A_2$ の方が高速であり，必ずしも導入したヒューリスティクスの効果が得られているわけではないことが分かる．その一方で，範囲に基づく高速化を用いた $A_2^R$ と，2 つの高速化手法を用いた $A_2^{GR}$ を比較すると，$A_2^{GR}$ の方が高速であり，組み合わせることにより，重心を考慮するの効果が得られることが分かる．

## 5.2 検索精度の評価

提案する枠組みにおける検索精度を検証するため，9 試合を対象に，サッカーにおける代表的なプレイ（オーバーラップやワンツーパスなど）を人手により抽出し，正解データセット $C$ を作成した．各正解データセット $C$ に対し，一つの要素 $q \in C$ をクエリとし，残りの $C \setminus \{q\}$ を正解集合とした場合の上位 $K$ 位検索を 10 回繰り返し，$K$-適合率のマイクロ平均を算出した．

クエリの次元数を $M$ とした場合，検索対象となる部分移動軌跡の総数はおよそ

$$25(Hz) \times 60(\text{秒}) \times 90(\text{分}) \times {}_{22(\text{人})}P_M \times 9(\text{試合})$$

と非常に大きくなるので，今回の実験では正解判定を緩く設定し，(1) 選手の組み合わせ $\pi$ が一致し，(2) 時間に重なりがある場合に正解であると判定する．具体的には，条件 $i-L+1 \leq j \leq i+L-1$ を満たす $D^{\pi}_{(j,L)}$ が正解集合に含まれている場合，$D^{\pi}_{(i,L)}$ を正解と判断する．実験結果を表 2 に示す．表中において，$C_1$ はオーバーラップに関するプレイ（正解数 : 218），$C_2$ は前方からの守備に関するプレイ（正解数 : 114），$C_3$ は後方からの守備に関するプレイ（正解数 : 309），$C_4$ はワンツーパスに関するプレイ（正解数 : 132）であり，クエリの次元数（人数）はすべて 2 である．

実験結果より，必ずしも高い精度が得られたわけではなく，移動軌跡だけではなく，ボールの位置や方向，他プレイヤとの距離など，外部の動的要因を考慮する必要があることが伺える．

# 6 まとめと今後の課題

本論文では，チームスポーツへの応用を念頭に，多次元移動軌跡データベースに対する上位 $K$ 部分移動軌跡検索問題を定式化するとともに，DTW 距離を用いた検索アルゴリズムを提案した．また，実際のサッカー選手の移動軌跡データを用いてその性能を評価した．

今後の課題としては，更なる高速化アルゴリズムの検討があげられる．また，動的属性 [2] の導入や時間差の考慮，選手の属性（チームやポジション）の反映なども重要な課題である．

# 参考文献

[1] L. Sha, P. Lucey, Y. Yue, P. Carr, C. Rohlf, and I. Matthews : Chalkboarding: A New Spatiotem-

[3] データスタジアム株式会社 (http://www.datastadium.co.jp) より

poral Query Paradigm for Sports Play Retrieval, *Proc. of the 21st International Conference on Intelligent User Interfaces*, pp.336-347 , 2016.

[2] H. Ohashi, T. Shimizu, and M. Yoshikawa : Flexible Similarity Search for Enriched Trajectories, *IEICE Transactions on Information and Systems* Vol. E100-D, No.9, pp.2081–2091, 2017.

[3] H. Sakoe and S. Chiba : Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol.26, No.1, pp.43–49, 1978.

[4] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh : Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping, *Proc. of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.262-270, 2012.

[5] S. Kim, S Park, and W. Chu : An index-based approach for similarity search supporting time warping in large sequence databases, *Proc. of the 17th International Conference on Data Engineering*, pp. 607–614, 2001.

[6] A. Fu, E. Keogh, L. Lau, C. Ratanamahatana, and R. Wong : Scaling and time warping in time series querying, *The International Journal on Very Large Data Bases*, Vol.17, No.4, pp.899–921, 2008.

[7] E. Keogh, L. Wei, X. Xi, M. Vlachos, S.H. Lee, and P. Protopapas : Supporting exact indexing of arbitrarily rotated shapes and periodic time series under Euclidean and warping distance measures, *The International Journal on Very Large Data Bases*, Vol.18, No.3, pp.611–630 , 2009.

[8] T. M. Rath and R. Manmatha : Lower-Bounding of Dynamic Time Warping Distances for Multivariate Time Series, Technical Report MM-40, Center for Intelligent Information Retrieval, University of Massachusetts Amherst, 2002.

原著論文 Original Paper

# ダンスの上手い人のマイニング的な分析

大北 剛 Tsuyoshi Okita
九州工業大学 Kyushu Institute of Technology
tsuyoshi.okita@gmail.com

井上 創造 Sozo Inoue
九州工業大学 Kyushu Institute of Technology
sozo@mns.kyutech.ac.jp

**keywords:** 行動認識, モービルコンピューティング, ユビキタスコンピューティング, 深層学習, ポーズ推定

**Summary**

IoT において,「歩く」,「立ち上がる」などの言語による行動のラベルを目的としたセンサからの行動認識を可視化する技術は,「歩く」,「立ち上がる」という言語による行動のラベルを目的とした映像からの行動認識との技術の融合を意味する．これは一転して,「歩く」,「立ち上がる」などの言語による行動のラベルのバイアスを排除する新たな行動認識の形を提案し，新たなマイニングのモデルを提案する. ダンスの上手い人と下手な人のどこが具体的に異なるかをセンサと映像からのマルチモダルな行動認識から探るプラットフォームの構築を報告する.

## 1. ま え が き

ダンスのような動作を伴う行動を複数の人間で比較する際にまずコーチなどの人間がいれば, ビデオを見て比較が可能となる. また, 三軸加速度などのセンサを用いてよい場合には, 何らかの動作を行なった場合のセンサ値を人間が比較箇所を特定して比較する方法が考えられる. 本論文においては, これを自動で行うにはどうすればよいかという問題を考えたい. まず, ダンスのような動作をビデオ、センサに記録する必要がある. そこで, 本論文ではセンサとビデオを記録するシステムを前半で開発する. 次に, それらの記録した情報をどう解析して人間が行うような解析に繋げるかの考察を行う.

これらの解析を行う動機の一つは行動認識の研究である. 人間の行動認識は, 大きくセンサベースの認識 [6, 5, 8, 7] とビデオベースの認識 (コンピュータビジョンではトラッキングともいわれる) のやり方に分類できる. このいずれのやり方においても, 各々のやり方で自然言語という形で表現された「歩く」「立ち上がる」などの行動へと分類する. この行動認識のやり方を用いると, A 氏が行ったダンスのある動作と B 氏が行った同じ動作を比較した場合に, これらが近い動作をしているか, かなり違う動作をしているかを判断することは比較的容易である. しかし, それらの動作が上手いか上手くないかという判断は極端に難しい. そこで, 設定を容易にするため, A 氏を上手い人と想定し, B 氏が A 氏の動作を真似る場合に, B 氏は A 氏の動作を上手く真似ているか否かという類似する問題へと摩り替えたい. この問題において, それらの動作が上手いか上手くないかという判断は可能となる.

本論文の貢献は以下の通りである.

- ビデオとセンサ信号を同一プラットフォームで収集するプラットフォームの開発,
- 取得したビデオを解析する仕組みとしてポーズ推定による骨組シークエンスを利用する方法の提案,
- ビデオとセンサ信号というマルチモダルなデータにおいてイミテーション学習の成功度を測定する方法として分散表現の累積和を用いる方法の提案

## 2. センサ/ビデオ取得システムの概要

### 2·1 ビデオ信号とセンサデータを収集するシステムの構築

まず, ビデオ信号とセンサデータを収集する必要があり, これは python ベースのシステムを開発した. 外観を図 2 に示す. ダンスを行なう人間側はスマートフォンを両手首 (2 台), 両足首 (2 台), 胸部 (1 台) の計 5 台装着すると想定し (図 1 に示す), これらのスマートフォンの三

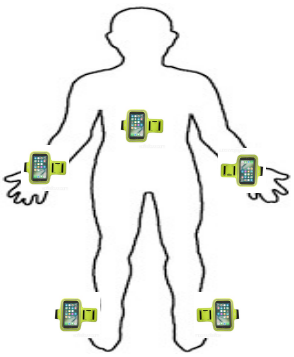


**図 1** 本論文において用いた 5 つのスマホの位置を示す.

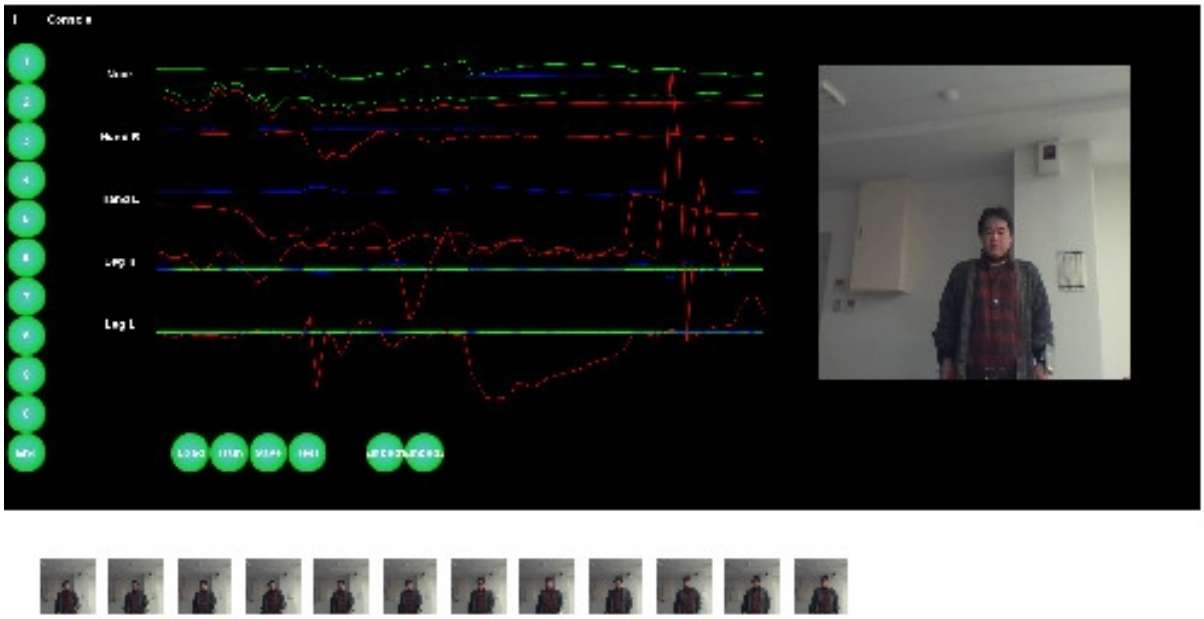


**図 2** センサ行動認識においては, 高次元のセンサの読取り値が時系列のシークエンスとして入力され, それに対応する動作を出力とする.

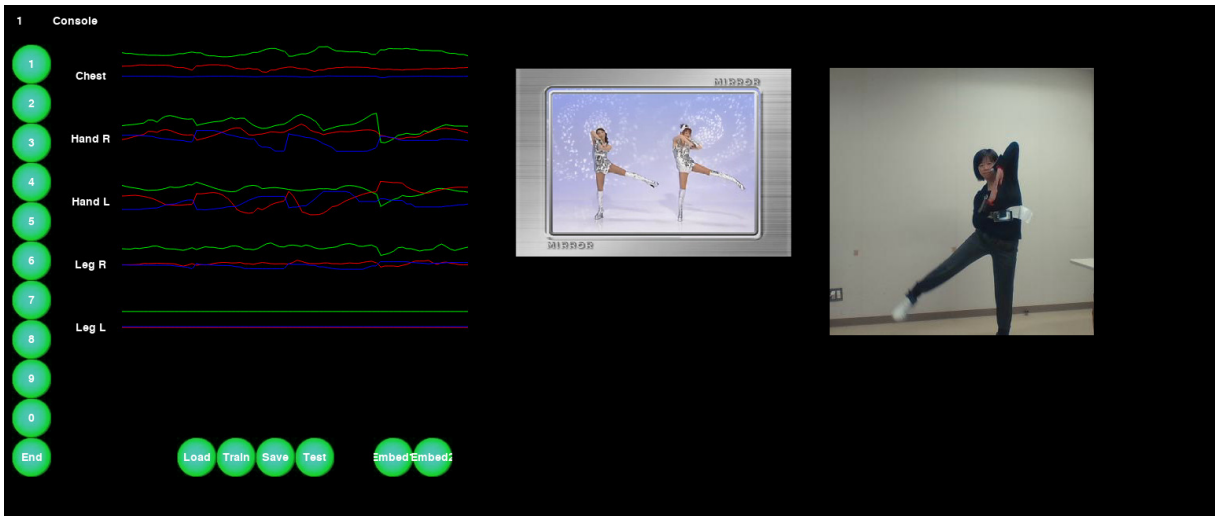



**図 3** 取得モードを示す. 中央に手本となるビデオ映像を流す.

軸加速度センサの値を本システムで記録する. 三軸加速度センサからの信号は UDP 経由で別々のチャネルで送信させ, 本システム側でその信号を受信する. 本システムはノートパソコン上に実装し, このためビデオ信号はウェブカメラから来ることを想定し, これを本システムで記録する.

内部構成は, センサ処理部ビデオ処理部からなる. センサ処理部においては, 時系列のセンサデータを取得し, 時系列に表示する機能を備える. ビデオ処理部においては, ウェブカメラからのデータを opencv を用いて取得して, システム画面に表示する機能を備える.

取得モード, 再生モード, 解析モードが存在する. 取得モードは図 4 に示す. このモードにおいては, センサデータとビデオ信号を記録する処理を行なう. 振付つきのダンスの取得モードにおいては, 手本となるビデオ画像を音声を伴って流し, これにより被験者が踊りやすくしている.

再生モードにおいては, 記録した信号を再生する. このモードにおいては, 手本となるビデオ画像は省略し, 記録したビデオ映像とセンサデータを再生することを可能としている.

解析モードにおいては, 本論文において述べるような骨組対骨組のアラインメント, 骨組のシークエンスの出力などの機能を持たせている.

### 2·2 骨組対骨組のアラインメント

取得したデータから骨組を得るためにビデオ映像を画像シークエンスに落とし, Openpose[1] を利用して各画像に対して骨組および 45 次元の骨組ベクトルの座標を得た. 図 5 はサムネイルの例を示し, また, 図 6 は骨組ベク

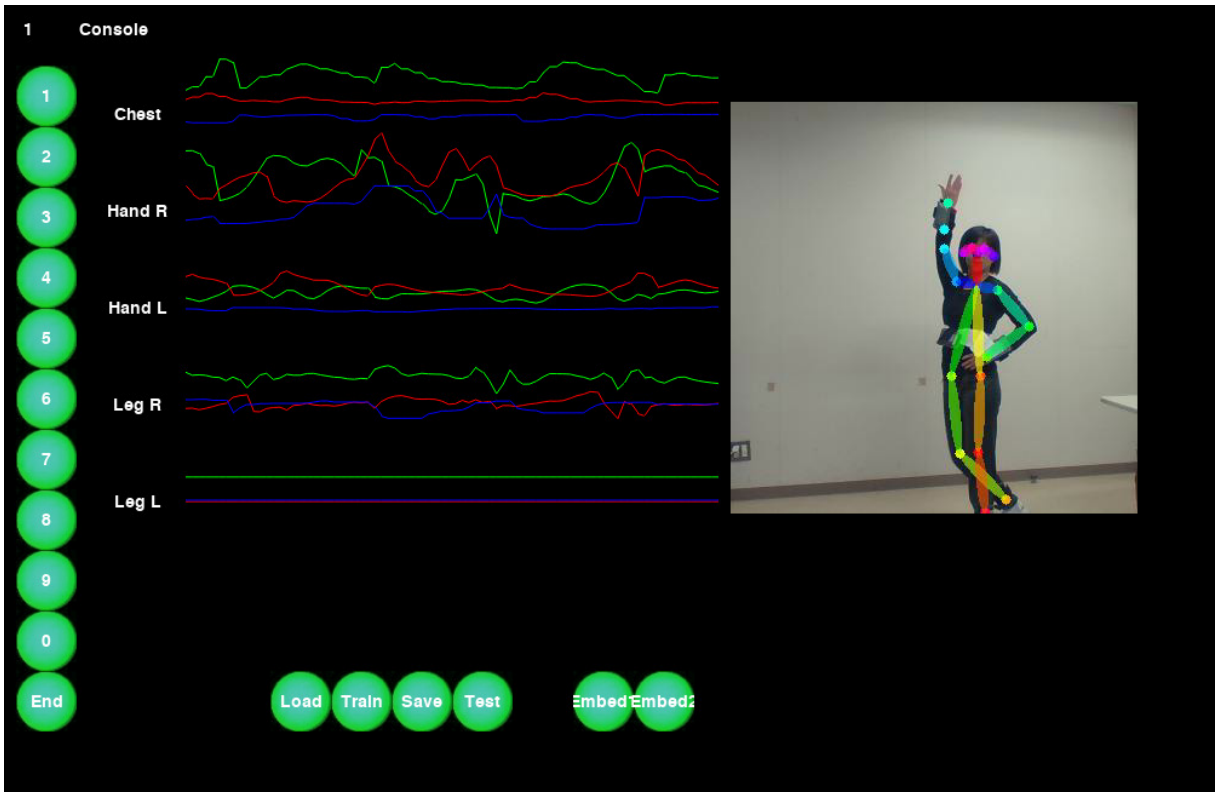



**図 4** 再生モードを示す. 左側には取得したセンサデータ, 右側には骨組解析をオーバーラップさせたビデオ映像を流す.

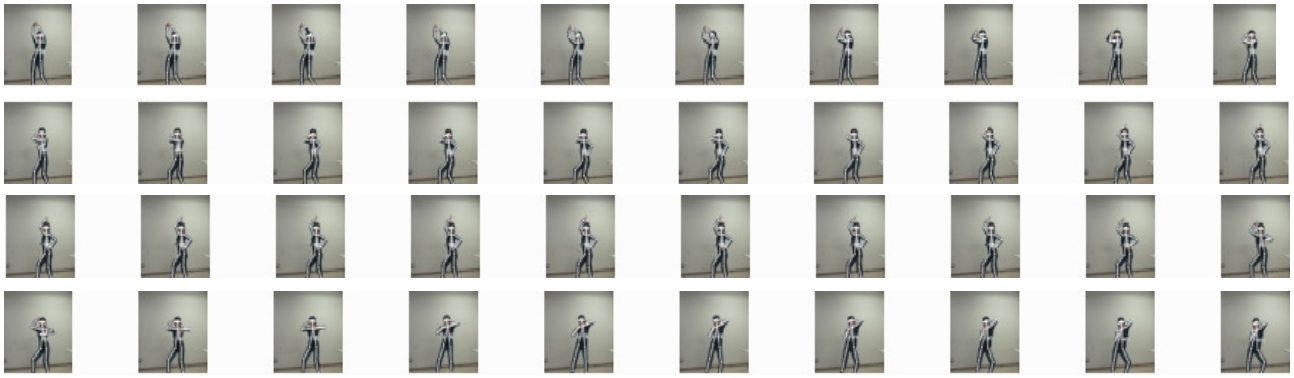


**図 5** 画像シークエンスにおいて, 各々の画像に Openpose を適用して, 骨組をそれぞれ得ている. この後さらにこれらの画像シークエンスをビデオ映像に変換する.

トルを表示したものを示す. これらの画像シークエンスをビデオ映像に変換した. ビデオ映像に変換した後はビデオ映像は骨組のみとなる.

## 3. ダンスの解析

ダンスを解析することは, 本論文においてはセンサデータおよび骨組ベクトルを解析することと帰着させる. 本論文においては, 手法を紹介して簡単な分析を行ない, 実行可能性を吟味するに留める. 自由なダンス曲で個人技の巧さを分析する形のものは考えず、固定したダンス曲、たとえばピンクレディの UFO の振付け, を上手い人と上手くなろうと努力している人の違いは何かを分析する形のものを考える. なお, 自由なダンス曲で個人技の巧さを分析する形においてはこのような形で上手い人のダンスの分析をする場合には, 使っているボキャブラリの数などを比較することができるはずである. この場合, 本論文の解析方法そのものでは対処はできず若干の拡張を必要とする.

主要なプロトコルは以下の 5 つとした.

(1) 上手い被験者 (以後, A と呼ぶ) がセンサをつけた状態でダンスを行ない, 同時にビデオ撮影も行なう. その後, 上手くなろうとする被験者 (以後, B と呼ぶ) がセンサをつけた状態でダンスを行ない, 同時にビデオ撮影も行なう.

(2) 振付けにおいてシーンを設定し, ビデオをシーン毎に区切るラベリングを行なう. この後, それぞれのシーンの開始を A と B においてアラインメントを

**図 6** 骨組シークエンスのみとする.

行なう. さらに, それぞれのシーンにおいて, A と B の骨組をアラインメントさせる.

(3) B が外して振付と全く違う動作を行なうことが頻出することが予想されるが, この場合には, アラインメントはあえてしないようにする.

(4) センサデータに関しては, ビデオにおけるシーンの区切りと同期する時点をそれぞれのシーンの開始点と定義する.

(5) この同期する点は, ビデオ信号とセンサデータの遅れを反映させた形のものとなり, しかし, すべての同期点において一定の時間の遅れと考える.

項目の 1 番目であるが, 我々の分析の目的を細分化して, イミテーション学習を行うコンテクストとする. つまり, 上手い被験者 (以後, A と呼ぶ) がセンサをつけた状態でダンスを行ない, 上手くなろうとする被験者 (以後, B と呼ぶ) がセンサをつけた状態でダンスを行う. このようにして記録したビデオとセンサデータを比較するという方法を取る.

前述したようにビデオは, ポーズ分析 [1] を行ない, 抽象化されて骨組となる. この抽象化により, 該当する部位, たとえば A の右手と B の右手, を比較することにより行なえることとなる.*1 また, 今回はこれ以外の比較は行なわず, 骨組のみで比較を行なう. 比較を行なうためには, シーン毎に A と B の骨組がアラインメントされている必要がある. このため, 項目の 2 番目であるが, A と B のアラインメントした骨組を比較するためには事前設定として, 振付けのシーンを設定して, ビデオ先導でシーンを区切っておく. つまり, それぞれのシーンの開始を A と B においてアラインメントを行なう. さらに, それぞれのシーンにおいて, A と B の骨組をアラインメントさせる.

項目の 3 番目は, A と B の距離を求めるのが困難な場合である. B が外して振付と全く違う動作を行なえば, A と B の距離の解釈が著しく困難となる. この場合には, アラインメントはあえてしないようにする. また, この期間は B のイミテーションを行なうスコアはゼロと考える. 一方, アラインメントしている場合には, A と B の骨組の差を比較して, そのスコアを骨組の要素ごとに行なう.

項目の 4 番目は, センサデータに関しては, ビデオにおけるシーンの区切りと同期する時点をそれぞれのシーンの開始点と定義するというものである.

項目の 5 番目は, センサデータとビデオのシーンとの間には遅れが存在する点で, これが無視できない. これは, 我々のシステムにおいて, センサデータは常にビデオ信号より若干遅れて到着し, この記録も若干遅れることによる. そこで, センサデータとビデオのシーンとの間の遅れはすべての同期点において一定の時間の遅れと考えるというものである.

## 4. 行動認識との比較

センサベースの行動認識は, 入力を高次元のセンサ信号とし, 出力を人間の行動とする. 教師あり行動認識として機械学習を用いる方法を本論文では論ずるため, ここでは教師あり行動認識のみを考慮する. 教師あり行動認識の場合, 行動クラス $\mathcal{Y}$ はトレーニング集合に有限個の行動として定義され, たとえば, $Y = \{$ 立ち上がる, 歩く, ジョギングする, $\ldots\}$ などとなる. 一方, センサデータ $\mathcal{S}$ は採取に用いるセンサの種類と数に依存し, センサデータの次元は $n(=\{1,\ldots,N\})$ となる. 各々のセンサは時系列のデータで構成され, $k$ 番目のセンサ $(1 \leq k \leq N)$ に対する時刻を $t(=\{1,\ldots,t_k\})$ と定義すると, $s^{(k)} = (s_1^k, s_2^k, \ldots, s_{t_k}^k)$ と表現できる. ここで人間の行動を記述するのは「立ち上がる」「歩く」「ジョギングする」などの自然言語である.

ビデオベースの行動認識は, 入力をビデオ信号とし, 出力を人間の行動とする. 本論文で行なうシーンの解析においては, 上手い人の行動をいかに上手く真似たか (イミ

*1 一方, この単純化は多くの現実的な因子を失うことは確かである. たとえば, 骨組のみをビデオにした場合, ダンスが上手いか下手かの判断が極端に難しくなる. 肉付けしても未だ難しく, テクスチャマッピングまでやらないと判断が難しくなるように思える.

テーションしたか) という因子を解析することになる. ここでも人間の行動を記述するのは「立ち上がる」「歩く」「ジョギングする」などの自然言語となる.

さて, 本論文における解析は, センサベースの行動認識やビデオベースの行動認識で用いるラベルの部分が大きく異なる. 一つ目, 我々の目的において, シーンに対してラベルづけは行なう. シーンはたとえば,「リズムキープ」「フォーコーナー」「腕をのばしたまま振る」「腕を止める」「ポーズ」「ウォーク」[*2]などの他, どちらかというとこれらが同じシーンにおいて A と B の動きを比較するための前準備に使いたい. マイニングできるためにはこれらの分類は必要となるがこれが本質ではない. 二つ目, これらの分類は上手く真似たかの指標とはなりえないため, 上手く真似たかの指標となるものを得る必要がある. 上手く真似たかの指標は, A と B のセンサデータとビデオ映像を比較する折に得られる. 三つ目, B の動作はエキスパートである A の動作とは似ても似つかぬ動作を行なっている可能性があり, これらの動作をしている箇所は比較の対象から外すのが無難であろうと思われる. このために, 同一のシーンにおいて, A と B の分散表現の距離の累積和を比較して, たとえばコサイン距離の累積和が閾値より大きければ対象から外すというやり方を選択した. 四つ目, センサデータの方も分散表現の距離の累積和の比較というやり方を選択した. 五つ目, 累積和の比較と言ったが, 振付の場合, 音楽と同期させるため, 同期ポイントはより明確なはずで同期した時刻における比較を行ない, かつ, 同期していない時刻における比較を行なわないのがよいかと思われる.

### 4·1 分散表現の距離の累積和の比較

ビデオ側においてもセンサ側においても分散表現を用いて比較するため, 以下のような方法を用いた. シークエンシャルな性質を考慮すると, センサ信号とビデオ信号/骨組の分散表現は等価である. したがって, センサ信号をエンコーダーデコーダ [2, 4, 9] を用いて翻訳して構築した分散表現を取り出して用いる方法である. そして, それをシーンにおいて A と B の分散表現の距離の累積和を比較するやり方である.

## 5. 実　　験

以上の動機のもと, センサの時系列シークエンスから行動のシークエンスへエンコーダーデコーダー型 [2, 4] を用いて翻訳することを考える. したがって, デコーダ側は時系列ごとにビデオシークエンスをラベルとして骨組ベクトルを用いる. この設定は率直にはいかないため, 以下のような工夫を行なう.

一つ目, 生のセンサ信号、ビデオシークエンスはいずれもストリームだが、エンコーダーデコーダー型では無限長を扱ってはおらず, 逆にセンテンス長の長い領域では精度が悪くなる. この長さ制限を考慮して、暫定的にストリームをエンコーダーデコーダーいずれの側においても, 100 ユニット以内に留めた. エンコーダー側はセンサの周波数、デコーダー側はビデオの周波数でいずれも異なるため、エンコーダー側の開始時点の時刻、終了時点の時刻に合わせる形でビデオ信号の始点、終点を定める. この制約を満たすように設定すると, エンコーダ長 61 項目, デコーダー長 45 項目となった. 測定に用いたセンサの周波数、ビデオの周波数は異なる被験者においても同じ設定で用いたため、この 61 項目と 45 項目という比率は固定とする.

二つ目, エンコーダ側の 1 項目は 3 軸加速度センサを 5 台装着したことによる 15 次元からなり, デコーダー側の 1 項目はビデオ信号をポーズ分析した結果の 45 次元からなる. つまり, 加速度センサの 15 次元, ポーズ情報の 45 次元を入力と考え, 一方, エンコーダ長, デコーダ長の 61 と 45 という項目数は, センサの周波数とビデオの周波数を修正したものと考えたことになる. なお, 加速度センサの 15 次元 (もしくはポーズ情報の 45 次元) を一般的に扱うことはせず, 凝集型 (agglomerative) なクラスタリングを行ない, ユークリッド距離に関して類似する点を同等と考え, 1000 クラスから 4000 クラスのラベルを貼った.

三つ目, 被験者は 3 人のデータを用いたが, 暫定的に個人差はないものと考え, また, ダンス, 運動などによる差はないものと考えた. 具体的には, (1) 歩く、走るなどの日常動作のデータ, (2) ラジオ体操第 1 のデータ, (3)UFO のデータの 3 種類を用いることにした. これら 3 種のデータは 20,000 対作成し, これをトレーニング, 検証, テストに 20,000 対, 700 対, 660 対にそれぞれ分割した.

この設定により, センサ信号の入力列があれば, 対応するポーズの列が出力されることになるがここではこの機能は使わない.

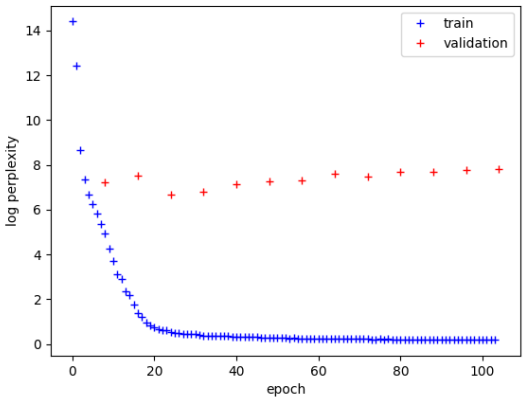



**図 7** パープレキシティの推移.

*2 これらの記述は http://www.asakumamasaru.com/entry/waack-dance による.

## 6. ま　と　め

本論文においては, ダンスの分析を行なうプラットフォームの構築と暫定的な解析を行なった.

問題が難しいため, 今後の話題はいくつも見付かった. 一つ目, 骨組ベクトルの表現を用いたが, 本当にこの表現で巧さの指標となることはもしかすると疑問かもしれない. なぜなら, これを映像にしてみた場合に上手いか下手かの判断がつきにくいことによる. 肉付けして, テクスチャマッピングが必要かもしれない可能性はある. 二つ目, エンコーダデコーダ型の学習器でデコーダ側のセンテンス長に制約をつけることが困難そうなことである. 三つ目, 凝集型のクラスタリングを用いたが, そもそも, クラス数が 4000 程度の場合, 骨組の映像を見た場合にスムーズさがなく, さらなるクラス数が必要そうであることである.

**謝　　辞**

伊藤雅子さん, 武田紳吾くんには貴重な時間を割いてダンスをしていただき感謝の意を表します.

## ◇ 参　考　文　献 ◇

Zhe Cao, Tomas Simon, Shih-En Wei and Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, In Proceedings of CVPR 2017, 2017.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. Conference on Empirical Methods in Natural Language Processing (EMNL

井上創造, ウェアラブルセンサを用いたヒューマンセンシング, 知能と情報, 28:6 pp. 170-186, 日本知能情報ファジィ学会, 2016 P 2014). 2014

Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate. ICLR 2015. 2015

井上創造, ウェアラブルセンサを用いたヒューマンセンシング, 知能と情報, 28:6 pp. 170-186, 日本知能情報ファジィ学会, 2016

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

Tsuyoshi Okita, Sozo Inoue. Recognition of Multiple Overlapping Activities Using Compositional CNN-LSTM Model. Ubicomp Poster, Sep, 2017.

Francisco Javier Ordonez, Daniel Roggen. “Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. S ensors 16:115, 2016.

Felix Hill, Roi Reichart and Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. Computational Linguistics. Vol. 41, No. 4, Pages 665-695. 2015.

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-04

# 段階的なテーマ詳細化によるニュース情報の体系的獲得

# Acquiring Organized Information from News by Incremental Theme Refinements

谷口 祐太郎[1*]　小林 哲則[1]　林 良彦[1]
Yutaro Taniguchi[1]　Tetsunori Kobayashi[1]　Yoshihiko Hayashi[1]

[1] 早稲田大学 理工学術院
[1] School of Science and Engineering, Waseda University

**Abstract:**
We propose an interactive system which allows a user to efficiently acquire organized information from a news corpus. Initiated by a topic specified by a user, the system first extracts a relevant set of sentences, from the corpus and divides them into themes each designated by an additional query term. The user then can select one of them, and the system maintains an interactive information access session by incrementally refining the search queries. This incremental process enables user's topic-specific yet efficient information access activities. Small-scaled experiments confirmed that the proposed system could provide the user with more effective supports than an ordinary search engine in the sense that it concisely provides a set of chronologically-sorted sentences classified by the themes. The experiments however further identified important issues, including the reduction of duplicate themes and sentences, and the deletion of sentences that cannot tell their meaning without neighboring contexts. For future work, we first implement automatic title generation, and then complete a total system.

## 1　はじめに

### 1.1　研究背景と目的

ニュースから体系的かつ効率的に情報を取得するために，ニュースコーパスからトピック（知りたい事柄）に関連する複数のテーマを抽出し，各テーマに関連する文が時系列順に並んだ文集合（ストーリーライン）を出力するシステムを提案する．

情報が個人や団体の資源になるといわれるなかで，世間にあふれている情報を得る労力を削減することは非常に重要である．とくに，ニュースの情報を得ることで世間の事件や事象を把握することは，日常会話の話題提示や，経済の未来予測，自らの意見の立ち位置の明確化に非常に有用である．ある事柄についてニュースから情報を取得しようとした場合，事柄に関するニュース記事群をそのまま読むと，事柄についての様々なテーマが混在しており量も多いため，非常に労力を要する．一方で，「Wikipedia」のような，ある事柄に関しての情報のまとめは，ニュースをもとに加工された，編集者の主観が含まれているような情報であり，客観性や公平性を望むことができないという問題がある．

したがって，ニュースから体系的かつ効率的に無加工の情報を獲得することは価値があると考える．

### 1.2　本研究の位置づけ

本稿では，ニュース記事集合とユーザが知りたい事柄を入力とし，テーマごとのタイトルつきの複数の文集合を出力とするシステムを提案するが，ニュースからの体系的・効率的な情報獲得の既存手法はさまざまである．

例えば，「ニュース検索結果を記事単位でクラスタリングした結果を見る」[2][3]，「ニュース全体をもとにテーマを抽出し，テーマごとの記事集合を図示する」[1] といった試みがある．また，「文書に対して，ユーザが指定したテーマに基づいて文を抽出する」[5] という試みを，ニュース記事全文をひとつの文書とみなしニュース記事に適用すれば，ニュースからの体系的・効率的な情報獲得につながる．これらの研究と本研究での提案手法を入出力の形式に着目してまとめたものを表 1 に示す．「タイトルを表す語」とは，「出力された各文集

*連絡先：早稲田大学 理工学術院
〒 162-0042 東京都新宿区早稲田町 27 早稲田大学 40 号館 701 号室
E-mail: taniguchi@pcl.cs.waseda.ac.jp

合または文書集合に，その内容を表すタイトルがつけられているかどうか」を示している．ここで，本研究と先行研究との差異に着目し，以下の 4 点に着目することで，提案システムを位置づける．

1. 入力にはユーザの意向を表す語が含まれる
2. 出力される集合の要素は文書ではなく文である
3. 出力される文集合または文書集合は複数ある
4. 各文集合または文書集合にタイトルを表す語がついている

**入力にはユーザの意向を表す語が含まれる**　ニュースから情報を取得するさい，ユーザが指定するトピックを入力として受け付けることで，ユーザの知りたい事柄を出力に反映させる．

**出力される集合の要素は文書ではなく文である**　ニュースからの情報取得の効率を考えると，適切な文集合が出力されるならば，文書集合より文集合を読むほうが好ましいと考えられる．あるトピックについての情報を取得するさい，同じ記事の中でもトピックに関連する文と関連しない文が存在し，必ずしも記事全文を読む必要はないため，文単位での集合を見るほうがよいとする．

**出力される文集合または文書集合は複数ある**　ニュース記事群の中には，トピックに関するさまざまなテーマが混在している．例えば，「オバマ」というトピックの「オバマの広島訪問」，「オバマの政策」，「オバマ政権への評価」といったテーマを別々に見ることで，ニュースを体系的に読むことが可能となる．

**各文集合または文書集合にタイトルを表す語がついている**　出力された複数の文集合または文書集合をユーザが実際に読むときには，ユーザは自分が読みたいと思った集合を選択するが，このときタイトルがつけられていれば，選択の支援となる．

# 2　提案システムとインターフェイス

## 2.1　システムの提案

ニュースから体系的かつ効率的に情報を取得するために，1.2 節で述べたとおり，「入力にはユーザの意向を表す語が含まれる」，「出力される集合の要素は文である」，「出力される文集合は複数ある」，「各文集合にタイトルを表す語がついている」という要素を満たすように，知りたい事柄を与えるとニュースコーパスから

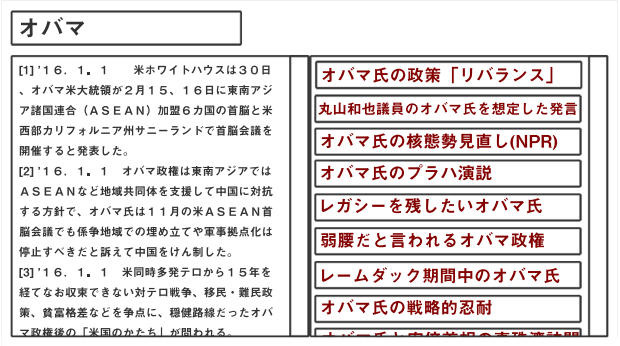



図 1: 想定インターフェイス (階層 1)

この事柄に関連するテーマを抽出し，テーマごとに関連する文集合を文単位で抽出し，時系列に沿って提示するシステムを提案する．

## 2.2　想定インターフェイスと必要事項

### 2.2.1　想定インターフェイス

提案システムを達成するためのインターフェイスを図 1，図 2 に示す．ここでは，知りたい事柄 (**トピック**) として「オバマ」が与えられた場合を例としている．以下に，想定する使用手順を示す．

1. 初期クエリとして，トピックである「オバマ」を与える．(図 1)
2. 左部分に，ニュースコーパスから検索された，「オバマ」に関連する文が時系列順に並んだ「**ストーリーライン**」が表示される．右部分に，「オバマ」に関連する文集合から取得された「**テーマ**」が表示され，各テーマにはタイトルがつけられている．(図 1) 本稿では，このタイトルを「**テーマタイトル**」を呼ぶ．
3. 右部分のテーマタイトルのうち一つ，例えば「オバマ氏のプラハ演説」を選択すると，左部分には，前段階から抽出された「オバマ」に関連する文集合からさらに抽出された，「オバマ氏のプラハ演説」に関連する文が，右部分にはその文集合からさらに取得されたテーマが表示される．(図 2)
4. ユーザは，興味のあるテーマを選択し，そのテーマに関連する文集合を読むことで，初期クエリに関する様々なテーマに基づいた情報を取得することができる．

表 1: 提案システムと先行研究の比較

| | 入力 | 出力 | | |
|---|---|---|---|---|
| | 形式 | 形式 | 複数であるか | タイトルを表す語 |
| 提案システム | 文書集合 + 語 | 文集合 | ○ | ○ |
| [2] | 文書集合 | 文書集合 | ○ | × |
| [1] | 文書集合 | 文書集合 | ○ | ○ |
| [5] | 文集合 + 語 | 文集合 | × | × |

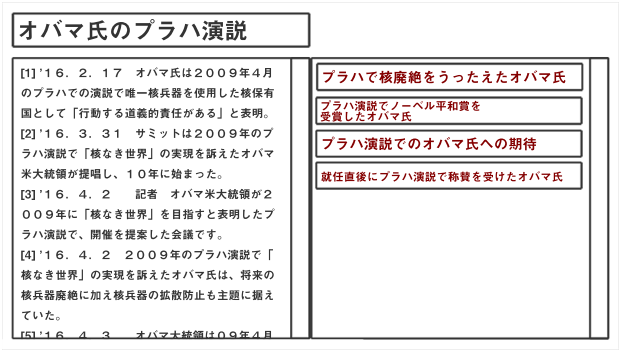



図 2: 想定インターフェイス (階層 2)

#### 2.2.2　想定インターフェイスの必要事項

2.2.1 節で提示したインターフェイスを達成するための必要事項を 3 点，以下に示す．

1. 適切なテーマ取得
   文集合を与えると，ある一つのテーマに沿った複数の文集合が導かれるような処理．
2. テーマタイトル生成
   ユーザがテーマを選択するきっかけとなるテーマタイトルの生成．
3. 前階層からの適切な文集合抽出
   各文集合についての，その文集合のテーマに関連する文の，前階層からの適切な抽出．

## 3　テーマ取得・文集合抽出

2.2.2 節で示した必要事項のうち,「(1) 適切なテーマ取得」と「(3) 前階層からの適切な文集合抽出」を達成するための手続きを検討し実装した．

### 3.1　テーマ取得・文集合抽出の提案手法

2.2.2 節で示した必要事項 (1) テーマ取得，必要事項 (3) 文集合抽出を実現するための手法を提案する．提案手法を図示したものを図 3 に示す．重要語を取得することで，適切なテーマ取得を行い，さらにその重要語を検索式に追加することで，適切な文集合を抽出するという狙いである．文の集合をもとに，重要語をつなげた検索式をユーザに提示する研究としては，松生ら [7] の研究が挙げられる．ただし，これは Web ページを対象として，ユーザが自身の検索する目的を明らかにするためにキーワード式を提示するという動機である．それに対して本研究は，ユーザが知りたい事柄に対して，様々な側面から情報を取得するということを狙っている．

本手法は，ユーザのテーマ選択に応じて文集合が生成されるのではなく，あらかじめ一定階層 d までの文集合を生成しておいて，ユーザの選択に応じてその文集合を表示するという想定である．擬似コード 1 で示したアルゴリズムによって，検索式とこの検索式により得られる文集合のタプルの集合を生成する．なお，擬似コード 2 のアルゴリズムを，擬似コード 1 で参照している．

知りたい事柄としてユーザが入力する初期クエリを q とする．

**Algorithm 1** テーマに沿った複数の文集合の抽出方法

**Input:** ニュース文集合 $\mathcal{S}$, 初期クエリ $q$, 階層数 $d$, 重要語数 $N$, 閾値 $t$(最少文数)

$First \leftarrow \{s | s$ は $\mathcal{S}$ のうち $q$ を含む文 $\}$
$OutSets \leftarrow \phi$
$NewSets \leftarrow \{(q, First)\}$
$OutSets \leftarrow OutSets \bigcup NewSet$
**for** from 0 to d-1 **do**
  $PreviousSets \leftarrow NewSets$
  $NewSets \leftarrow \phi$
  **for all** $set$ in $PreviousSets$ **do**
    $Temp \leftarrow Algorithm2(set[0], set[1], N, t)$
    $NewSets \leftarrow NewSets \bigcup Temp$
  **end for**
  $OutSets \leftarrow OutSets \bigcup NewSets$
**end for**

**Output:** $OutSets$

**Algorithm 2** 検索式と文集合のセット (Q, S) からのテーマに沿った複数の文集合の抽出方法

**Input:** 検索式 $Q$，文集合 $S$，重要語数 $N$，閾値 $t$(最少文数)

$OutSets \leftarrow \phi$
$K \leftarrow \{k|k$ は $S$ に含まれる単語のうち 1式で示した手法で重要語抽出したときの上位 $N$ 語 $\}$
**for all** $k$ in $K$ **do**
  $Temp \leftarrow \{s|s$ は $S$ のうち $k$ を含む文 $\}$
  **if** $|Temp| \geq t$ **then**
    $NewQ \leftarrow Q \bigcup k$
    $TempSet \leftarrow (NewQ, Temp)$
    $OutSets \leftarrow OutSets \bigcup TempSet$
  **end if**
**end for**

**Output:** $OutSets$

1. ニュース文集合から，q を含む文を抽出し，文集合 S(q) とする．

2. q を含む文集合に含まれる単語を，重要度が大きい順に $N$ 語抽出する．この重要語を $(x_1, x_2, \cdots, x_N)$ とする．なお，重要度の指標は，式 (1) を用いた．

$$R_{morph} = \frac{c_{morph}}{c_{all}} \cdot \frac{C_{all}}{C_{morph}} \tag{1}$$

ただし，式 (1) の $R_{morph}$ は，単語 morph の重要度，$c_{morph}$ は文集合における morph の出現回数，$c_{all}$ は文集合における全単語の出現回数，$C_{morph}$ はコーパス全体における morph の出現回数，$C_{all}$ はコーパス全体における全単語の出現回数を示す．文集合における morph の出現頻度を，コーパス全体における morph の出現頻度で正規化したといえる．コーパス全体における morph の出現頻度の log をとった式で重要度を測る場合もある [4] が，本研究に適用したところ，出現頻度が少ない単語が上位に来たため，log を使わない式を用いることとした．

重要語を抽出するさい，自立語のみという制約を設けた．また，単語をカウントするさい，上の階層で既に取得した検索式に含まれる任意のキーワード（初期クエリ含む）との距離（原文における単語間の距離）が 3 以下，上の階層で既に取得した検索式に含まれる任意のキーワード（初期クエリ含む）との間に助詞が含まれる，という場合のみカウントするようにした．これは，前段階のキーワードと同格にあるような語を重要語に含めないという狙いがある．例えば，「オバマ」に対して「大統領」というキーワードは，多く共起するが，文集合を限定する意味が少ない．また，「する」,「いる」,「こと」,「ない」は，自立語であるが意味の少ない語とみなし，stop words とした．

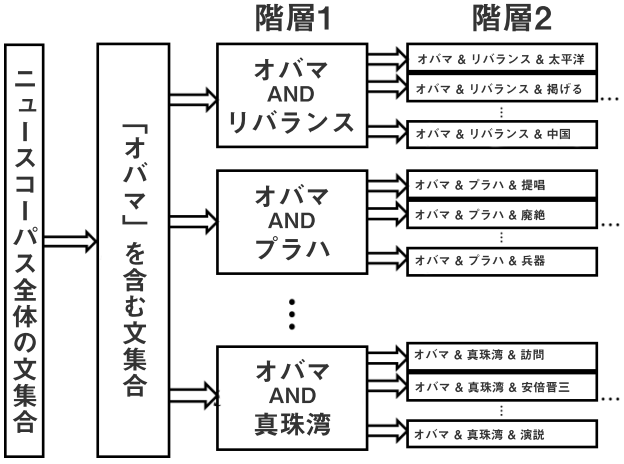



図 3: 文集合抽出の提案手法

3. 「q AND $x_n$」を含む文を抽出し，それぞれの文集合を S(q AND $x_n$) とする．これを 1≤n≤$N$ について行う．（階層 1 の文集合）

4. 「q AND $x_n$」の文集合に含まれる単語から，さらに重要語を抽出する．このとき，重要度が大きい順に $N$ 語抽出する．この重要語を $(x_{n1}, \cdots, x_{nN})$ とする．これを 1≤n≤$N$ について行う．

5. 「q AND $x_n$ AND $x_{nm}$」を含む文を抽出し，それぞれの文集合を S(q AND $x_n$ AND $x_{nm}$) とする．これを 1≤m≤$N$, 1≤n≤$N$ について行う．（階層 2 の文集合）

6. 以上を階層 d まで繰り返し，各検索式と検索式から導かれる文集合のタプルの集合を出力とする．ただし，文集合に含まれる文の数が t 未満のときは，出力に含めない．

## 3.2 テーマ取得・文集合抽出の実装

### 3.2.1 前処理

2016 年の毎日新聞の記事が収録されたコーパス「CD-毎日新聞 2016 データ集」の各記事データの「日付」と「本文」を利用して実験を行った．3.1 節で示した提案手法を実現するために，以下の項目が満たされるように前処理を行った．

1. 各記事データの本文が，文単位に分割されている．

2. 分割された各文が，単語単位に分割されている．また，各単語が自立語であるかどうかの情報を保持している．

**記事本文の文分割**　各記事に対して，記事本文を入力として，文単位に分割されたリストを出力とする。文分割は，以下のようなルールで行った．

1. ”。”，”！”，”？”のいずれかと，その次の文字の間で分割する．
2. ただし，カギ括弧内では分割しない．
3. パラグラフの末尾は必ず分割する．

**各文の単語分割**　各文に対して，単語単位に分割されたリストを出力とする．単語分割には，形態素解析システム Juman および構文解析システム KNP の固有表現を特定する機能を用いた．Juman の形態素解析による一形態素を一単語をみなす．ただし，人，場所の固有表現と解析された場合は，形態素としては分割されていても，一つの固有表現を一単語とみなす．(例:「安倍」「晋三」→「安倍晋三」) また，Juman の解析による自立語であるかどうかの情報を保持する．

#### 3.2.2　提案手法の実装

3.1 節で示した提案手法の実装について述べる．

初期クエリ q，最大階層 d，取得重要語数 N，文数閾値 t を入力として，階層数，検索式，文集合を値として持つ dict のリストを出力する．ここで，同階層で文集合が完全に一致するものはマージし，各検索式の最後のキーワードを両方保持する．例えば,「アベノミクス，税収，１億」を検索式にもつ文集合と「アベノミクス，税収，活躍」を検索式にもつ文集合が完全に一致した場合,「アベノミクス，税収，(１億，活躍)」を検索式にもつ文集合として出力する．また，類似度が95%以上である文のペアについて時系列が早いほうの文のみを結果に含めることで，文の重複を回避する．

# 4　提案システムの評価実験

提案システムとインターフェイスの評価実験を行う．[6] では，インターフェイスの必要事項のうち「文集合抽出」に着目し，文集合の正解データを作成することで，取得された各テーマに対してどれだけ正確に文が抽出されたかについての評価を行ったのに対し，本稿ではインターフェイスのモックアップを作成し被験者に利用させることで，提案システムとインターフェイス全体についての評価を行う．

## 4.1　実験手法

#### 4.1.1　提案システムの準備

2.2.2 節で示したインターフェイスの必要事項のうち,「テーマ取得」と「文集合抽出」を 3.1 節で示した手法で実装し,「タイトル生成」を手動で理想的なタイトルをつけることで実現する．あらかじめ定めたトピック (「オバマ」,「ＥＵ」,「東芝」) それぞれを入力として,「CD-毎日新聞 2016 データ集」をコーパスとしたときの出力結果から，インターフェイスのモックアップを自動生成する (提案システム)．このときのシステムの実行条件を，表 2 に示す．図 4 に,「オバマ」をトピックとするモックアップの画面を示す．

#### 4.1.2　ベースラインシステムの準備

オープンソースの全文検索サーバー「Fess」[1]を用いて,「CD-毎日新聞 2016 データ集」を自由なクエリで記事単位検索できるようにする (ベースラインシステム)．検索エンジンはニュースなどから情報を獲得する目的で日常的に利用されており，提案システムの比較対象として自然であると考えた．

#### 4.1.3　実験内容

提案システムとベースラインシステムをそれぞれ用いて，3 つのトピックに関する情報を調べて，トピックごとにまとめを作成することを 5 人の被験者に要求する．ここでいうまとめの作成とは，利用者が獲得した情報の中で有益だと思うものについて記述するという意味であり，高度な要約などは要求していない．まとめを作成するさいに，ニュース文からコピーアンドペーストすることを許す．

評価実験の公平性を保つために，被験者 A〜E のうち，被験者 A，C，E は,「オバマ」,「東芝」について調べるときは提案システムを先に利用し,「ＥＵ」について調べるときはベースラインシステムを先に利用するようにし，被験者 B，D は逆の順番に 2 つのシステムを利用するように実験を設計した．

実験中に，被験者は実験者に質問・意見・感想を自由に述べる．最後に被験者は,「ニュースから体系的に情報を獲得できたのはどちらのシステムか.」,「ニュースから効率的に情報を獲得できたのはどちらのシステムか.」という 2 つの質問に対して,「提案システム」,「どちらかというと提案システム」,「どちらともいえない」,「どちらかというとベースラインシステム」,「ベースラインシステム」の 5 つからそれぞれ解答を選択する．

[1] https://github.com/codelibs/fess/releases/tag/fess-11.4.9

表 2: モックアップ作成時のシステム実行条件

| | |
|---|---|
| 使用コーパス | 「CD-毎日新聞 2016 データ集」 |
| トピック | 「オバマ」,「ＥＵ」,「東芝」 |
| 取得重要語数 | 300 |
| 階層数 d | 3 |
| 文数閾値 t | 5 |

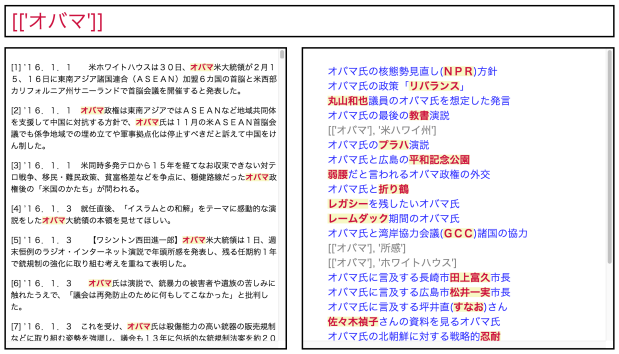



図 4: システムのモックアップ

## 4.2 実験結果と議論

### 4.2.1 トピックに関して作成されたまとめについて

図 5 に，被験者 A による「オバマ」に関するまとめ (ベースラインシステム利用)，図 6 に，被験者 A による「オバマ」に関するまとめ (提案システム利用) を示す．各被験者のまとめを比較すると，システムの特性上，ベースラインシステムよりも提案システムのほうが，まとめる観点が被験者によって一定している傾向にある．また，例えばベースラインシステムを利用した「オバマ」に関するまとめでは，5 人中 3 人の被験者が，オバマ氏の後任であるドナルド・トランプ氏について言及しているが，提案システムが提示するテーマにはトランプ氏に関するものは含まれていない．すなわち，提案システムには，多くの利用者にとって重要なテーマであっても表示されないものがあるという問題点があり，テーマ取得の方法 (重要語の指標) に改善が要求されていると考える．

### 4.2.2 質問回答結果について

各被験者は実験終了後に，「(質問 a) より体系的にニュースから情報を獲得できたのはどちらのシステムか.」,「(質問 b) より効率的にニュースから情報を獲得できたのはどちらのシステムか.」という 2 つの質問に対して,「1. ベースラインシステム」,「2. どちらかというとベースラインシステム」,「3. どちらともいえない」,「4. どちらかというと提案システム」,「5. 提案システム」の 5 つからそれぞれ解答を選択する．回答結果を表 3 に示す．これによると，5 人中 4 人が質問 a に対して 4 または 5 と回答し，5 人中 3 人が質問 b に対して 4 または 5 と回答しているため，提案システムはニュースから体系的かつ効率的に情報を取得することに貢献しているといえる．なお，被験者 A が質問 a に対して 1 と回答したのは,「提案システムは，提示されたテーマがトピックに関する情報を網羅している保証がないため，体系的とはいえない」との理由である．

表 3: 体系的・効率的であるかの質問に対する被験者の回答結果

| 被験者 | 回答 | |
|---|---|---|
| | 質問 a | 質問 b |
| A | 1 | 5 |
| B | 4 | 5 |
| C | 5 | 5 |
| D | 4 | 2 |
| E | 5 | 3 |

### 4.2.3 被験者の意見・感想

各被験者は，実験の最中に，自由に各システムに対する意見・感想を述べることができる．以下に，意見・感想から導くことができる提案システムの問題点を示す．

**提示されるテーマの問題**　1 人の被験者が，各トピックに関して提示されたテーマがトピックのすべての情報を網羅しているとは限らないということを指摘した．1 人の被験者が，提示されたテーマに偏りがあるのではないか，と述べた．1 人の被験者が，提示されたテーマがトピックに関して重要なテーマであることの保証がないということを指摘した．

**文集合内の文の冗長性**　2 人の被験者が，同じ文集合内に似たような文が含まれていて冗長に感じるという意見を述べた．3.2.2 節で述べた通り，重複度が 95%以上の文のペアは，一方を削除することで重複を解消しているが，重複度の閾値の決め方に改善が要求されていることを示す．

**文集合間の文の冗長性**　2 人の被験者が，異なるテーマの文集合であるにもかかわらず，含まれる文がほとんど同じであるため冗長に感じるという意見を述べた．3.2.2 節で述べた通り，含まれる文集合が完全に一致する文集合はマージするようにしているが，例えばある 2 つの文集合の Jaccard 係数が一定以上の場合はマー

ジするといったように，マージする条件の緩和が要求されていることを示す．

**テーマの階層数の設定が不十分である**　2 人の被験者が，さらに階層を進んで文を絞りたいと思ったときに，表示されるテーマがないという問題を指摘した．階層数や文数閾値の決め方に改善が要求されていることを示す．

**文単体では意味が通らないことがある**　3 人の被験者が，文集合に含まれる文の中には，主語が省略されているなどの理由から，文単体では意味がわからないものがあるという指摘をした．この問題の解決策の一つとして，インターフェイス上で文を選択すると，文を抽出する元となった記事を表示させるようにするというものが挙げられる．また，抽出された文の前後の文を，意味が通る最少の文数だけ表示させることで，効率性を維持することができると考える．

以上の問題点の他に，2 人の被験者が，1.2 節で述べたシステムの特色である「文単位で出力されること」に対して，有用であると言及した．3 人の被験者が,「テーマごとに出力されること」に対して，有用であると言及した．また，1 人の被験者が，2.2.2 節で述べたインターフェイスの必要事項である「タイトルがつけられていること」に対して，有用であると言及した．2 人の被験者が，文が時系列順に並んでいることに対して，有用であると言及した．

## 5　まとめ・今後の計画

知りたい事柄 (トピック) を与えるとニュースコーパスからこのトピックに関連するテーマを抽出し，関連する文集合を提示するシステムを提案し，これを達成するための必要事項 (「テーマ取得」「タイトル生成」「文集合抽出」) を提示した．そのうち「テーマ取得」「文集合抽出」に着目し，段階的に重要語を検索式に追加する手法によって達成を試みた．提案システムの評価手法を提案し，評価を行うことで，提案システムが目的を達成していることを示した．また，被験者からの意見・感想を求めることで，実際に利用者が感じる提案システムの有用性と問題点が明らかとなった．

今後は，本稿では手動で実現した「タイトル生成」を自動で行う手法を提案し，インターフェイスを全自動化する．実際に利用者が自由に検索できるように提案システムを実装する．また，実験により明らかになった問題点の改善手法を検討する．一方で，明確な評価基準が存在しないようなシステムに対してどのような評価手段がありうるかについての議論を深める．

## 参考文献

[1] Philippe Laban, Marti A. Hearst, ”newsLens: building and visualizing long-ranging news stories”, Proceedings of the Events and Stories in the News Workshop, pp. 1–9, 2017.

[2] Srinivas Vadrevu, Choon Hui Teo, Suju Rajan, Kunal Punera, Byron Dom, Alex Smola, Yi Chang, Zhaohui Zheng, ”Scalable Clustering of News Search Results”, Proceedings of the fourth ACM international conference on Web search and data mining, pp. 675–684, 2011.

[3] Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma, ”Learning to Cluster Web Search Results”, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 210–217, 2004.

[4] 大竹清敬，岡本大吾，児玉充，増山繁，”重要文抽出，自由作成要約に対応した新聞記事要約システム YELLOW”，情報処理学会論文誌，Vol.43，pp. 37–47，2002.

[5] 砂山渡，谷内田正彦，”観点に基づいて重要文を抽出する展望台システムとそのサーチエンジンへの実装”，人工知能学会論文誌，Vol.17，pp. 14–22，2002.

[6] 谷口祐太郎，小林哲則，林良彦，”ニュースからのトピックに関するストーリーラインの生成”，言語処理学会，第 24 回年次大会，2018.

[7] 松生泰典，是津耕司，小山聡，田中克己，”検索結果の概要を表すキーワード式生成による質問修正支援”，電子情報通信学会データ工学ワークショップ（DEW2005），1C–i9，2005.

- 自らの選挙戦で「チェンジ（変革)」、そして「イエス・ウィ・キャン（やればできる)」のスローガンを掲げて華々しく登場。２００９年１月、黒人初の米大統領に就任
- 就任間もなく「核兵器なき世界」の実現を訴えノーベル平和賞を受賞したが、外交、安保、内政とも理想と現実のはざまで揺れた２期８年だった。
- イラクとアフガニスタンでの二つの戦争の「終結」を公約に掲げ、２０１１年末にイラクからの軍撤退を果たしたものの、過激派組織「イスラム国」（ＩＳ）の台頭により、１４年には再派遣に追い込まれ、戦争終結は実現できなかった
- 5/27, 原爆投下国のトップとして初めて被爆地・広島を訪問し、平和記念公園（広島市中区）での演説で改めて「核兵器なき世界」を訴えた
- 12/9 の会見で, 大統領選への干渉を狙ったサイバー攻撃について再調査するよう情報機関に指示
- イラク戦争終結や医療保険制度改革（オバマケア）など多くの公約を達成
- 核開発問題を巡り対立していたイランとは１５年７月、英仏独中露の５カ国と共に包括的共同行動計画を結んだ。イランは核兵器取得につながる核計画の規模縮小に応じ、米国など主要国は制裁解除で応えた。キューバとの国交回復も果たした。
- 就任４カ月前に発生したリーマン・ショックで米経済は深刻な金融危機に陥っていたが、米史上最大規模の景気対策に着手。
- イラク駐留米軍撤退という最大の公約を果たしながら、過激派組織「イスラム国」（ＩＳ）の増勢を許し、２０１４年に再度イラクへの軍事介入を余儀なくされた
- 大統領は現実的でなく脇が甘い」など批判的な意見

図 5: 被験者 A による「オバマ」に関するまとめ (ベースラインシステム利用)

- ＴＰＰを政権の「遺産（レガシー)」にしたい
-「雇用を生み、環境にやさしい経済を打ち立てるため、革新の精神を育てないといけない」と強調
- プラハ演説で「核兵器なき世界」を提唱
- 広島市の平和記念公園で改めて、「核兵器なき世界」の理想を追い求める決心を強調
- 停滞する核軍縮機運を再び高める
- 任期の最後に「レガシー（政治的な遺産)」を築く狙いがある
- 国内では医療保険制度改革（オバマケア)、リーマン・ショックで落ち込んだ米国経済の回復に取り組んだことを挙げた
- 経済、軍事分野で台頭する中国を念頭にアジア太平洋地域を重視する「リバランス（再均衡)」政策を進め、ＡＳＥＡＮとの連携を重要な柱の一つとしている
- 過激派組織「イスラム国」（ＩＳ）や国際テロ組織アルカイダの脅威が続く状況で大統領選の年を迎えたため、一般教書演説で、「米国民を守り、テロ組織と戦うことが最優先事項だ」と明言
- 弱腰政権と批判された

図 6: 被験者 A による「オバマ」に関するまとめ (提案システム利用)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-05

# ブログテキストの分析に基づく語の意味の経時変化可視化の試み

# An Experimental Result on Visualization of Word Sense Changes by Blog Text Analysis

石川 雅弘 [1*]
Masahiro ISHIKAWA[1]

[1] 高崎健康福祉大学
[1] Takasaki University of Health and Welfare

**Abstract:** More than a decade have passed since blog or SNS became common. Massive amount of user-generated text has already been accumulated on the web. Many researchers are trying to analyze accumulated texts trying to exploit them in many fields. In such analysis, treatment of text meaning is important. Text is composed of words, thus word sense treatment is essential. However, word meanings undergo changes, thus we should consider word sense changes over time. In this paper, we present a method to detect and visualize word sense changes. The proposed method uses Random Indexing technique, which is based on the distributional hypothesis of word meanings. The result of an experiment on blog texts is also presented.

## 1 はじめに

Web や SNS，レビューサイトなどの普及により，記者や作家などの職業的文章生産者だけではなく，一般個人により生産された大量のテキストの蓄積が進んだ．そこにはかつてならば音声発話として消えていったような個人的な意見や感情の表明も含まれており，流行分析や評判分析，マーケット分析など様々な活用が試みられている [1]．今後もテキストデータの蓄積は継続的に拡大していくと考えられ，その有効活用を考える必要がある．

自然言語で記述されたテキストの活用において重要な課題の一つが意味処理である．そこでは，テキストの文字列としての一致・不一致だけではなく，その表す意味を適切に扱う必要があるが，その基礎として，単語の意味の扱いが重要である．

しかし，単語の意味は時間とともに変化する可能性がある．変化の速い現代においては単語の意味変化や新たな語義の獲得も速いと考えられるが，職業記者のように統制されていない一般個人の言語使用においては，その傾向は一層強いであろう．また，例えばトレンド分析など経時変化に対する感度が重要な分析においては，その考慮が一層重要である．企業や商品に対するイメージや人気の変化も，企業名や商品名の利用文脈の変化としても表れると考えられ，単語の意味・利用文脈の変化を分析することはマーケティングなどにおいても有用性があると考えられる．

このような観点から，ブログデータを対象として，単語の意味変化の検出とその可視化を試みた．本稿ではその手法と結果を報告し，今後の課題を検討する．

*連絡先：高崎健康福祉大学 健康福祉学部 医療情報学科
〒 370-0033 群馬県高崎市中大類町 37-1
E-mail: ishikawa@takasaki-u.ac.jp

## 2 単語のベクトル表現

自然言語処理おいては，単語や文書を数値ベクトルとして表現するベクトル空間モデルが一般的である [2]．基本的なベクトル空間モデルでは，$n$ 個の文書の集合を $m \times n$ 単語–文書行列 $\mathbf{C}$ で表現する．

$$\mathbf{C} = \begin{pmatrix} c_{11} & \cdots & c_{1j} & \cdots & c_{1n} \\ \vdots & \ddots & & & \vdots \\ c_{i1} & & c_{ij} & & c_{in} \\ \vdots & & & \ddots & \vdots \\ c_{m1} & \cdots & c_{mi} & \cdots & c_{mn} \end{pmatrix}$$

$\mathbf{C}$ の要素 $c_{ij}$ は，単語 $i$ の文書 $j$ における出現頻度や TF-IDF 値などの重みであり，列ベクトルが文書を，行ベクトルが単語を表す．また，単語や文書間の（非）類似度は，ベクトル間のユークリッド距離やコサイン尺度などで表わされる．このように一つの単語や一つの文書が一つの次元を構成する 1-of-k スタイルの表現では，行列 $\mathbf{C}$ は巨大な疎行列となり効率的な処理が難しい．また意味的処理にも適さないことから，LSI(Latent

Semantic Indexing)[3] などの手法により，より低次元で密な行列に変換されることが多い．しかし，全文書を入手してから一括して処理をする必要があるなど，時間を追って単語の意味変化を分析するためのベクトル表現方法としては適さない．

word2vec[4] や Random Indexing[5] では，巨大な疎行列を介さずに最初から密でより低次元な単語の分散表現を生成できる．word2vec は生成した単語ベクトル間に加法構成性があると見られるという点でも注目されているが，Random Indexing はより単純な計算で単語ベクトルを生成できる上，文書集合が増加した時の漸増的計算も容易であるという利点があり，時間とともに変化する単語ベクトルの生成手法として適している．また，できるだけ大規模で網羅的な処理を行う場合でも，処理の分散と結果の集約が容易という利点がある．そのため，本研究では Random Indexing を用いて単語ベクトルを生成し，それを分析することで意味の変化の検出と可視化を試みる．

### 2.1 Random Indexing

Random Indexing は，単語の意味の分布仮説に基づく単語ベクトル生成手法である [5]．分布仮説では，ある単語の意味はその出現文脈に現れる他の単語群により決定されるとされる．例として次のような文章を考える（ここでは分かち書きした各部を単語とする）．

春 は 桜 が 咲き ます

ここで各語には「索引ベクトル」と呼ばれる固有のベクトルが割り当てられているとする．この時，各語の前後 $k$ 語の範囲をその語の文脈とし，文脈中の各語の索引ベクトルを合計することでその単語のこの出現における「文脈ベクトル」を得る．文脈ベクトルは単語のその文脈による意味付けである．例えば単語の前後 2 語までを文脈とすると,「春」,「は」,「桜」,「が」,「咲き」の各索引ベクトルの合計が「桜」のこの出現における文脈ベクトルである．ある単語のベクトル表現は，その単語の全文書における全出現の文脈ベクトルを合計することで得られる．

単語に索引ベクトルを割り当てる時点では単語の意味も類似性も不明のため，索引ベクトルは互いに直交であることが望ましい．しかし，1-of-k スタイルで各語に互いに直交なベクトルを割り当てると，単語の異なり総数に等しい $m$ 次元が必要となり，疎な高次元ベクトルとなってしまう．

Random Indexing では，索引ベクトルの次元を単語の異なり総数 $m$ より小さな値 $m'$（$\ll m$）とし，各語の索引ベクトルとして $m'$ 次元の擬直交ベクトルを割り当てる事で $m'$ 次元の単語ベクトルを生成する．これは高次元ベクトル空間では次元数より遥かに多い擬直交ベクトルが存在するという性質を利用している．ここで $\vec{u},\vec{v}$ が擬直交ベクトルであるとは，$\vec{u}\cdot\vec{v}\approx 0$ を意味する．

なお，一定の条件を満たしたランダムなベクトルを生成することで擬直交ベクトル群を得られることが示されており [6]，本研究では論文 [7] で提案された手法を用いる．

Random Indexing では，単語ベクトルは文書集合全体におけるその単語の全出現の文脈ベクトルを単純に合計することで生成できる．そのため，時間とともに文書集合が増加する場合でも，逐次的に新たな文書における文脈ベクトルを求め，それを過去の文書集合から計算された単語ベクトルと合計することで最新の単語ベクトルを求められる．したがって，新たな文書における文脈ベクトルとそれを合計した最新の単語ベクトルが時間とともにどのように変化するかを分析することで，単語の意味変化を捉え得ると考えられる．

## 3 提案手法

### 3.1 単語ベクトルの生成法

本稿で分析対象とするのは，Web から収集したブログ記事テキストである．ブログ記事には作成日時が付されているため，記事集合を一定期間ごとに分割し時間順に整列することができる．

対象とするブログ記事の集合を $D$ とし，それらを月別に分割し $D_0, D_1, \cdots, D_{T-1}$ とする．$D_t$ は第 $t$ 月目に生産されたブログ記事集合である．

文書集合から Random Indexing に基づいて単語ベクトルを生成するには，単語の文脈を前後何単語の範囲とするかを決める必要があるが，今回は簡単のため単語が含まれるブログ記事テキスト全体を文脈とする．すなわち単語 $w_i$ の $D_t$ から求めた単語ベクトルは

$$v_i^{(t)} = \sum_{d\in D_t}\sum_{w\in d} w \text{ の索引ベクトル}$$

となり，これを単語 $w_i$ の第 $t$ 月における月別ベクトルと呼ぶ．また，第 $t$ 月目までの全てのブログ記事から得られる $w_i$ の単語ベクトルは

$$V_i^{(t)} = \sum_{j=0}^{t} v_i^{(j)} = \sum_{j=0}^{t}\sum_{d\in D_j}\sum_{w\in d} w \text{ の索引ベクトル}$$

であり，これを単語の第 $t$ 月における累積月別ベクトルと呼ぶこととする．第 $t$ 月における $w_i$ の単語ベクトルは，累積月別ベクトルを長さ 1 に正規化したものであり，下式で与えられる．

$$\hat{V}_i^{(t)} = \frac{V_i^{(t)}}{\|V_i^{(t)}\|}$$

本稿の目的は，単語ベクトルの変化を追跡することで，分布仮説に基づく単語の意味，すなわち利用文脈の変化を分析し，その可視化を試みることである．

### 3.2 変化の追跡と検出方法

ベクトル表現された単語間の類似度としては，一般的に用いられているコサイン尺度を採用する．長さが 1 に正規化された二つの単語ベクトル $\vec{u}, \vec{v}$ のコサイン尺度は下式で与えられる．

$$cosine(\vec{u}, \vec{v}) = \sum_i u_i v_i$$

単語の意味が恒常的であれば，二つの期間 $s, t$ における単語ベクトル $\hat{V}_i^{(s)}, \hat{V}_i^{(t)}$ はほぼ等しいことが期待でき，類似度は

$$cosine(\hat{V}_i^{(s)}, \hat{V}_i^{(t)}) \approx 1.0$$

となる．逆に意味に変化があれば類似度は低下する．したがって，各時点の単語ベクトルについて，同じ単語の過去の時点でのベクトルとの類似度を追跡することで意味変化を検出できる可能性がある．

### 3.3 変化内容の分析方法

単語ベクトルは，文脈中に共起した単語集合の索引ベクトルを合算したものであり，単語ベクトルの変化は共起する単語集合が変化したことを意味する．したがって，共起する単語集合のクラスター構造の変化を分析することで，意味変化の内容を知ることができると考えられる．

単語ベクトルのクラスタリングには，SOM (Self-Organizing Maps) [8] の一種である Batch Map を用いる．SOM はデータを二次元空間や一次元空間上に整列したセルに写像し，データ空間上のクラスター構造を低次元空間上に「再現」する．そのため高次元空間中のクラスター構造の可視化に利用される．本研究では，SOM によるクラスタリング結果の可視化手法としては [10] で提案した極座標ヒストグラムと面グラフを修正した手法を用いる．これらについては 4 で述べる．

本研究で用いる Batch Map はバッチ学習型の SOM であり，一般的な逐次学習型の SOM と比べて計算効率が良い．また，後述する近傍半径が 0 の場合には k-means クラスタリングと一致する．また，一般的な SOM や k-means クラスタリングではデータ間の非類似度としてユークリッド距離を用いるが，ここでは類似度としてコサイン尺度を用いる．従って，本研究で用いる SOM は Dot Product Batch Map[8] であり，近傍半径が 0 の場合 spherical k-means クラスタリングに一致する．データの写像先のセルは k-means クラスタリングのクラスターに対応するが，SOM ではそれらの間に二次元または一次元上の隣接関係が与えられ，隣接したセルには類似したクラスターが配置されるようにクラスタリングが行われる．そのため，得られたクラスター間の類似の度合いを測ることができる．本研究では、環状に配置されたセルを用いる．

#### 3.3.1 Dot Product Batch Map

Dot Product Batch Map によるクラスタリング手順を示す．ここで，セルの数は k-means クラスタリングにおける $k$ であり，求めるクラスター数に対応する．また，セルは環状につながっているものとする．

1. 各セルの中心ベクトルを初期化する．
2. 各データをコサイン尺度が最も小さいセルに割り当てる．
3. 各セルの中心ベクトルを，近傍半径 $R$ 内のセルに割り当てられた全てのデータの平均ベクトルで更新する．
4. 収束するまで (2), (3) を繰り返す．ただし手順 (3) の半径 $R$ は大きな値から徐々に減少させる．

## 4 実験

本節では，実際のブログデータを対象とした単語ベクトルの変化と変化内容の可視化例を示す．ただし，追跡期間において明らかに使用文脈に変化が生じたと考えられる単語である「福島」のみを対象とした．この単語の使用文脈の分析は，社会が受けたインパクトの大きさや風評の広がりと収束の様子を知るためにも有意義だと考える．

### 4.1 データセット

2011 年から 2012 年にかけて goo ブログ [9] の新着記事 RSS で捕捉した 34756 ブロガーのうち，2011 年より前に「福島」を含む記事を投稿しており，かつ 2012 年 3 月 11 日以降も記事を投稿している 5714 ブロガーが 2010 年 1 月 1 日から 2012 年 3 月 31 日までに投稿した記事を対象とした．

記事テキストは MeCab(v0.97)[11] により形態素解析を行ない，名詞と判定された語のみを抽出し分析に用いた．ただし代名詞，非自立語，接頭辞，接尾辞，数詞，サ変接続詞は除いた．また，出現数が 10 未満のものと 1 文字のみのものも除いた．なお，MeCab 用辞書としては IPA 辞書を用いた．

最終的に対象となったブログ記事数は 3,690,657，単語の異なり総数は 507,532 である．

また，月ごとの変化を追跡するために，2010 年 1 月から 2012 年 3 月までのひと月ごとにデータセットを分割した．

## 4.2 単語ベクトルの生成

3.1 で示した手順に従い，全ての単語の全期間の月別単語ベクトルと累積単語ベクトルを作成した．索引ベクトルは 200 次元であり，したがって単語ベクトルも 200 次元である．

## 4.3 自己類似度の変化の可視化

まず,「福島」ベクトルの第 $t-1$ 月と第 $t$ 月の月別ベクトル間類似度，第 0 月と第 $t$ 月の累積月別ベクトル間類似度を求める．結果を図 1 に示す．横軸は 2010 年 1 月を第 0 月とした月数である．一見して第 14 月，すなわち 2011 年 3 月に，それ以前とも以後とも大きく異なる月別ベクトルが生成されていることが分かる．また，それに伴い累積ベクトルにも急激な変化が生じている．

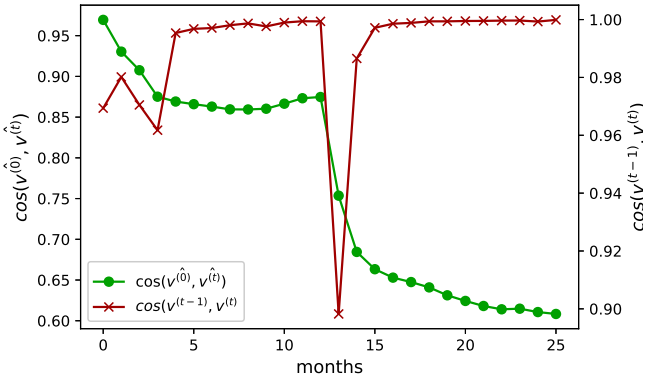



図 1: 「福島」ベクトルの自己類似度の変化

比較のため，同時期に「福島」ほどの変化は生じなかったと考えられる「沖縄」の自己類似度の様子を図 2 に示す．第 14 月に「福島」のような変化は見られないことが分かる．

## 4.4 極座標ヒストグラムによるクラスター構造の可視化

図 1 から，第 14 月に「福島」の出現文脈が大きく変化したことは分かった．ここからはその変化の内容を分析する．

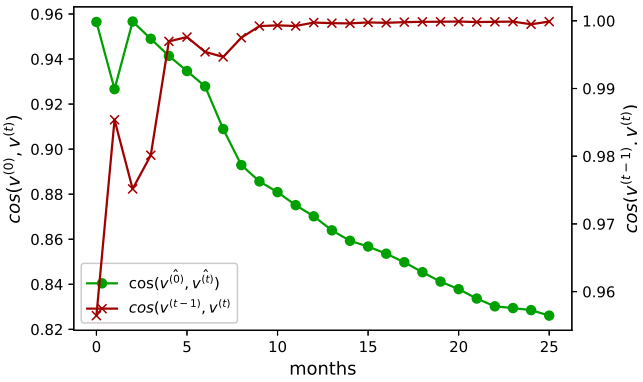



図 2: 「沖縄」ベクトルの自己類似度の変化

文脈の変化とはすなわち文脈内で共起する単語集合の変化である．そこでまず，最終月における全ての単語ベクトルをクラスタリングした結果を図 3 に示す．SOM は，k-means クラスタリング同様クラスター数 ($k$) を指定する必要があるが，本実験では $k = 128$ とした．図の各バーが一つのクラスターを表し，バーの長さはそのクラスターを構成する単語の数を表す．隣接した二つのバーの間の角度は，二つのクラスターの中心の分離度を示す．すなわち，角度が小さいほど二つのクラスターの中心同士の類似度が大きく，狭い角度に多くのクラスターが密集している範囲は，実際には大きな一つのクラスターが，大きする $k$ を与えられたため過剰に分割された可能性がある．また，隣接した二つのバーの角度が小さいほど類似した色が配されている．

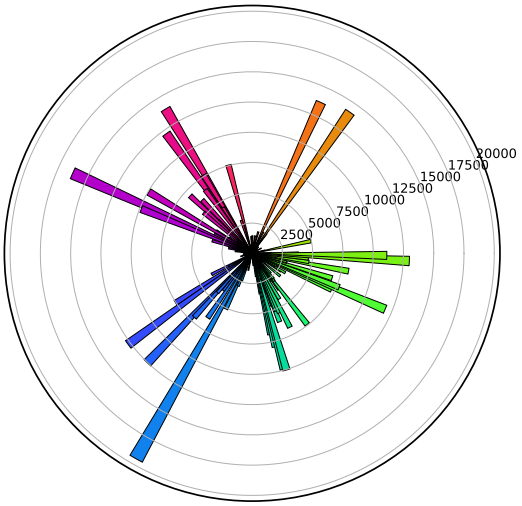



図 3: 最終月における全単語ベクトルのクラスター構造（バーの長さはクラスターに属する単語の数）

この図からは，全単語群が全体的にいくつかの大きなクラスターに別れ，さらに大クラスターが細分され

ているらしい様子が分かる．

### 4.5　文脈とその変化の可視化

図 3 のクラスター構造をベースに,「福島」の文脈の変化の可視化を試みる．

図 4 に示すのは，2011 年 2 月と 3 月それぞれについて,「福島」と共起した単語のみに絞って描いた極座標ヒストグラムである．ただしバーの長さは，共起した単語のうちそのクラスターに属する単語がその月に出現した頻度である．彩色については図 3 とは異なる方針を取り，隣り合ったクラスターがなるべく異なる色を持つようにした．これは，密集した領域でクラスターサイズの変化があった場合，類似色では判別しづらいからである．

このヒストグラムは，いわば「福島」のその月の文脈を可視化したものであり，異なる期間のヒストグラムを比較することで，その変化を見ることができる．図中矢印で示したように，2 月には大変小さかった二つのクラスターが 3 月には爆発的に増大していることがわかる．図 1 の自己類似度チャートの第 14 月における大きな変化はこのクラスターが原因だと考えられる．

### 4.6　面ヒートマップによる文脈変化の可視化

図 5 に示すのは，図 4 と同じデータの全期間分を一枚の面グラフとして表示したものである．横軸が月数，縦軸は 128 個のクラスターの相対サイズを積み上げたものである．また，各期間の各クラスターの相対サイズを表す領域は相対サイズの値でヒートマップ風に彩色した．青がサイズが小さいことを，赤が大きいことを表している．

面グラフは，全期間に渡る全クラスターの変化の様子を一枚で可視化できる点が優れているが，面積だけで表しているとクラスター数が多い場合には判別し難いという問題があったが，ヒートマップ風彩色により変化を把握しやすくなっている．

図中楕円で示したように，第 14 週において二つのクラスターが増大していることがわかる．これは図 4 で増大していた二つのクラスターと同じものである．

これらのクラスターを精査することで，実際にどのように単語の使用文脈が変化したのかを分析できる．

## 5　まとめと今後の課題

ブログや SNS などの一般ユーザーが生産した大量のテキストデータの蓄積を背景に，テキストデータの利活用が進んでいる．しかしテキストデータの活用のためにはまずは単語の意味を適切に扱う必要がある．しかし単語の意味そのものも時間とともに変化するため，変化があった時にそれを捉えられる必要がある．

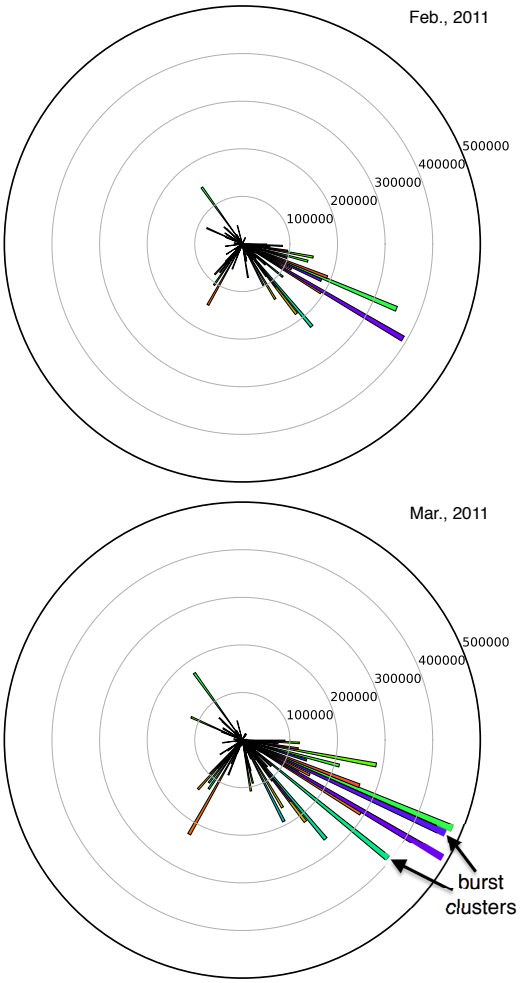



図 4: 2011 年 2 月・3 月にそれぞれ「福島」と共起した単語群のクラスター構造（バーの高さはクラスターに属する単語のその期間における出現頻度）

そこで分布仮説に基づいた単語ベクトル生成手法である Random Indexing を利用して，単語の意味の経時変化の可視化を試みた．対象としたデータセットの期間内において大きな変化があったと考えらえた「福島」を例として，自己類似度チャート，Batch Map による単語クラスタリング，極座標ヒストグラムと面ヒートマップによる可視化例を示した．

例えば商品名をターゲットにこのような分析を行えば，商品の人気や評判がどのタイミングでどのように変化したのかを知るのにも役立つと考えられる．

しかし今回分析の対象としたのは，極めて例外的に短時間で大きな文脈変化があったと考えられた単語であり，その他の一般的な単語の意味変化を捉え可視化できるかどうかは明らかではない．その確認のために

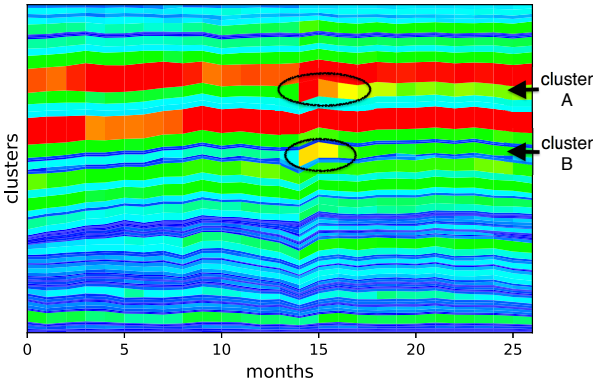



図 5: 「福島」と共起した単語群のクラスター構造の面グラフ・ヒートマップ表現（幅と色は相対クラスターサイズを表す）

は，データセットの量，特に期間の延長が必要であり，より長期にわたってより穏やかに変化する単語の分析を行う必要がある．今回使用したブログデータをベースに，今後それ以降の記事の収集を進めたい．

また，今回のように特定の単語をターゲットに据えて分析するのではなく，変化のあった可能性のある単語そのものを発見するには，ヒストグラムやヒートマップはあまり役立たないであろう．単語の数は膨大であり，その全てのチャートを作成して目視することはできないからである．変化のあった可能性のある単語を自動的に検出し，それらについてのみチャートを作成し目視するのが現実的である．

したがって，変化を自動検知する手法を考える必要があるが，これは各単語の自己類似度を追跡することで可能であると考えており，今後取り組みたい．

極座標ヒストグラムや面ヒートマップの基礎となるクラスタリングにも問題が残る．

今回の実験では最終期間における累積ベクトルを用いてクラスタリングを行い，過去の各期間のヒストグラムやヒートマップはそのクラスターに各期間の単語を割り当てて作成した．しかし，単語ベクトルは変化しているというのが本研究の出発点である．したがって，最終期間の累積ベクトルと，過去の時点での「同じ」単語の累積ベクトルは異なっている可能性がある．表層の文字列が同じだからと言う理由で異なる期間の単語を「同じ」ものとして扱うことはできないはずである．しかしそれらを別のものとして扱うためには，表層文字列とは別の「概念」を表す記号が必要となり，煩雑となる．また，現在扱っているデータセットの期間は短いため，このような配慮が実際に必要になる単語はほとんどないと考えられ，試みとしての本実験では追求しなかった．将来的により長期に渡る意味変化の追跡を行う中ではこの点も考慮したい．

## 参考文献

[1] Junichi Kato, *Customers' Needs for Digital Terrestrial Television Broadcasting: An Analysis of Weblog Data*, Proceedings of The 8th International Conference on Innovation and Management, pp.1093–1096, 2011.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, 1999.

[3] Christos H. Papadimitriou et al., *Latent Semantic Indexing: A Probabilistic Analysis*, In Proc.of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp159–168, 1998.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, *Distributed representations of words and phrases and their compositionality*, Advances in Neural Information Processing Systems 26, pp3111-3119, Curran Associates, Inc. 2013.

[5] Sahlgren, Magnus, *An Introduction to Random Indexing*, In Proc. of the Methods and Applicatons of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, 2005.

[6] Kanerva P, Kristofersson J, Holst A, *Random indexing of text samples for latent semantic analysis*, In Proc. of the 22nd Annual Conference of the Cognitive Science Society, p.1036, 2000.

[7] Dimitris Achlioptas, *Database-friendly Random Projections*, In Proc. of the 20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp274–281, 2001.

[8] Teuvo Kohonen, *Self-Organizing Maps, Third Edition*, Springer-Verlag, 2001.

[9] goo ブログ, `http://blog.goo.ne.jp/`.

[10] Masahiro Ishikawa, *Visualizing Cluster Structures and Their Changes over Time by Two-Step Application of Self-Organizing Maps*, Proceedings of the 2011 International Workshop on Behavior Informatics at the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011), pp.160–171, Shenzhen, China, May 2011.

[11] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-06

# 着目点の明示によるデータ分析支援

# Data Analysis Support by Displaying Points to Notice

中川拓郎　砂山渡　畑中裕司　小郷原一智
Takuro Nakagawa　Wataru Sunayama　Yuji Hatanaka　Kazunori Ogohara

滋賀県立大学 工学部
School of Engineering, The University of Shiga Prefecture

**Abstract:**
現在，爆発的に増大しているデータを分析し価値のある知識を発掘するニーズが高まっている．しかし，膨大で捉え所のないデータの分析に際して，目をつけるべき着目点を見出すことは簡単ではない．そこで本研究では，データの中から分析の手がかりとなる着目点をハイライトにより明示をすることで，データ分析の支援を行う枠組みを提案する．データ分析において平均等の基準値からのズレが大きいデータを着目点として明示するとともに，与えた着目点を手がかりとして，データへの着目と絞り込みを繰り返し行える機能を実装した．着目点の明示機能を利用することで，データ分析の支援が可能であるかを評価実験を行い，提案システムの推奨する着目点の明示機能を利用することで，分析者のデータ解釈の数が増加する傾向がみられた．

## 1　緒論

現在，爆発的に増大しているデータを分析し価値のあるデータを発掘するニーズが高まっている [1]．しかし，膨大で捉え所のないデータから，目をつけるべき着目点を発見することは簡単ではない現状がある．

この問題点に関して，多量のデータの中から着目点を探す事が困難であるデータマイニングに慣れていない人に対して基本的な着目点を見出す支援を，データ分析に慣れている人には，データの見落としを最小限にする幅広い着目点を見出す支援がそれぞれ必要である．

そこで本研究では，テキストマイニングのための統合環境 TETDM （Total Environment for Text Data Mining）[2] をベースとして，データから平均や頻度の基準値からのズレが大きいデータを着目点として明示するとともに，与えた着目点を手がかりとして，データへの着目と絞り込みを繰り返し行える機能を実装することにより，初心者でも分析を行いやすいデータ分析環境の構築を行う．

## 2　関連研究

### 2.1　データ分析の着目点に関する研究

レビュアがレビューの評価値を決定した根拠となる商品の機能や特徴の提示を行う研究がある [3]．この研究では，レビュー文章中の商品に関する文章を抽出し，使われている形容詞について着目することで，日本語評価極性辞書からレビュー評価の根拠を示している．本研究では，レビュー全体から評価値や単語でハイライトと絞り込みを行うことでレビュー評価の根拠を解釈できる着目点の明示を行う点で異なる．

アンケートデータの解析時に他の人の回答との関連度が低い少数回答を明示する研究研究がある [4]．この研究では少数回答，少数意見に重点をおいて着目点を示しているが本研究では，さまざまな目的に対応して幅広く着目点になり得るデータを明示を行う点で異なる．

### 2.2　データの可視化に関する研究

テキストマイニングによる授業評価アンケートの分析時に共起ネットワークを用いた自由記述の可視化を行う研究がある [5]．この研究では，アンケートの自由記述から取り出した頻出語の 30 単語を用いて共起ネットワークによる可視化を行っている．本研究ではデータ中から分析者が示した着目点を直接ハイライトする可視化を行う点で異なる．

## 3　データ分析における着目点の明示機能

### 3.1　データ分析の流れ

図 1 にデータ分析の流れを示す．まず，分析を行うデータを入力する．入力した膨大な量のデータは一度

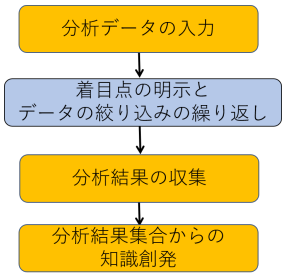



図 1: データ分析の流れ

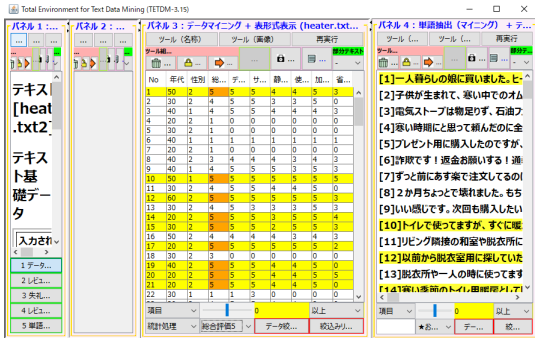


図 2: TETDM の画面の例

にすべてのデータを理解することができない．そのため，分析者が条件を選択しデータの明示，絞り込みを繰り返し行い，データを少なくすることで分析結果を見出す．分析結果集合からデータを整理，統合することで知識創発を行う．

データ分析を行うためには、データの絞り込みと繰り返しによってたくさんの分析結果を集めることが重要である．

## 3.2 データ分析環境：TETDM

データ分析環境の基盤として，テキストデータマイニングのための統合環境 TETDM[1]を利用した．

TETDM は，ツールを独自に開発，追加できる特徴を持つデータ分析の統合環境である．テキスト分析のための多様なツールを有し，テキストデータと数値データの両方を扱える．

TETDM の画面例を図 2 に示す．TETDM の 1 画面は，複数のパネルから構成されており，各パネルにデータ処理のためのツールと，データ可視化のためのツールをペアでセットして利用する．TETDM には，40 以上の処理ツールと可視化ツールが実装されており，それらを柔軟に組み合わせて利用することができる．

[1]http://tetdm.jp からダウンロード可能

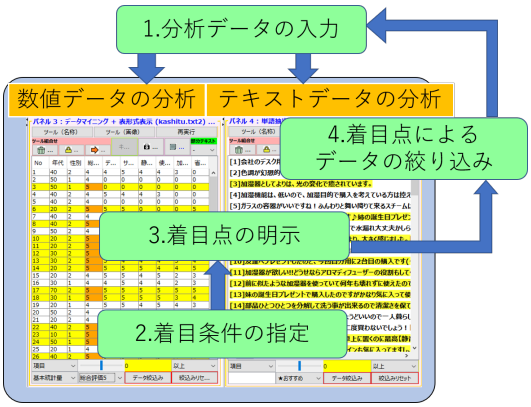



図 3: 着目点の明示とデータの絞り込みの流れ

## 3.3 データ分析における着目点の明示機能の枠組み

松本ら [6] により，TETDM を用いたデータ分析時に，数値分析ツールとテキスト分析ツールを連携して利用する環境が提案されている．しかし，データ分析時にどこに着目点してデータを分析するかという点について，分析者が自発的に着目点を発見する必要があった．

本研究では，何らかの条件に合致する一部のデータに対して着目点を明示し，データの絞り込みを繰り返し行う機能をデータ分析ツールとテキスト分析ツールに追加することで，分析者がデータ中の傾向や特徴を見出す支援を行う．

図 3 に，本研究で TETDM に追加した着目点の明示とデータの絞込み機能の流れを示す．すなわち，入力された分析データに対して，着目条件を設定し，その条件にマッチするデータを明示する．また，明示したデータをより深く分析するために，データの絞り込みを行って分析を繰り返す．この着目条件を指定する際に，条件の設定を手動で行いづらいという問題点を解決するために，おすすめの着目条件を提示する．以下の節で，この各ステップについての詳細を述べる．

## 3.4 分析データの入力

入力データは，数値データとテキストデータの両方，またはいずれか一方のみのデータセットを入力とする．各データは，属性と属性値のペアで構成されている．

TETDM においてテキストは以下のように分割して処理される．すなわち，テキストデータ全体を「文章」，テキストデータに挿入される「スナリバラフト」というタグで区切られた部分テキストを「セグメント」，句点で区切られた文を「文」として処理する．入力データの例を表 1 に示す．この例のようなデータにおいて

表 1: 本分析環境の入力データの例

| ID | 年代 | 性別 | 総合得点 | テキスト |
|---|---|---|---|---|
| 1 | 50 | 2 | 5 | 一人暮らしの娘に買いました。ヒーターは小さくて軽いので片手で運べるしちゃんと暖まれるので良いです！w だそうです。 |
| 2 | 40 | 2 | 3 | 2か月ちょっとで壊れました。もちろん保証で新しい物と交換できました。 |
| 3 | 40 | 1 | 4 | いい感じです。次回も購入したいと思います。 |
| 4 | 50 | 1 | 5 | トイレで使ってますが、すぐに暖かです。。 |
| 5 | 60 | 2 | 5 | 以前から脱衣室用に探していたところ、娘の家で使っていて快適だとのお墨付きをもらい購入しました。大満足です。 |
| 6 | 30 | 2 | 4 | 脱衣所や一人の時に使ってます。センサーがついてるのでつけっぱなしにならないので、電気代の節約になってると思います。デザインもシンプルで気に入っています。値段もお手頃でした。 |
| 7 | 30 | 2 | 5 | 夜中の授乳用に購入。すぐに温風が出るので短時間であたたかくなるので良かったです。 |

は，一人分のデータごとに「スナリバラフト」を挿入して，1 つのセグメントとして扱われるように前処理を行う．

## 3.5 着目条件の指定方法

### 3.5.1 数値データ分析ツールの着目条件

TETDM 上で数値データの明示と絞り込みを行うために用意した処理ツール「データマイニング」を図 4 に示す．

数値データ分析ツール「データマイニング」による着目条件の設定方法には次の 2 種類がある．

1 分析者による着目条件の設定

2 おすすめによる着目条件

1 の分析者による着目条件の設定は，以下の手順で行う．

1) 着目する属性を選択する．

2) スライダーで属性値の閾値を決定する．

3) 閾値以上か閾値以下を選択する．

すなわち，存在する各数値データの中から，1 つの属性を選択して，その属性値の範囲を選択して指定する．明確な条件設定が分析者の頭の中にある場合は，これを用いることで柔軟な条件設定が可能となる．

しかし，どのような条件を入力すべきかわからない分析の初心者や，一通り思いつく条件を設定した後に条件を模索したい場合などに，簡易に条件を入力できるおすすめ着目条件を，表 2 のように用意する．

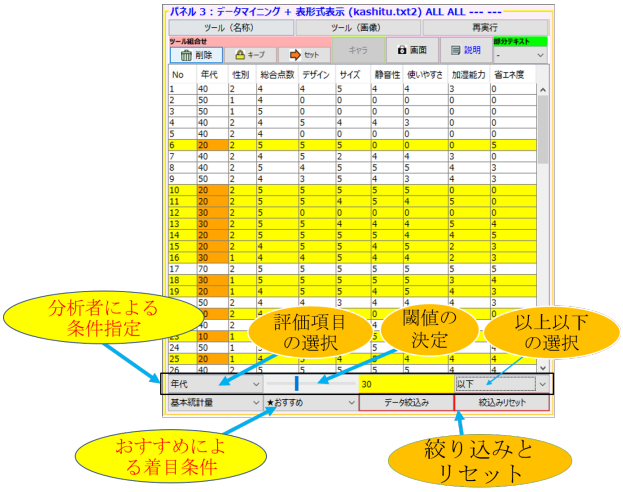



図 4: ツール「データマイニング」の操作画面

表 2: 数値データ分析ツールのおすすめ着目条件

| 条件名 | 条件内容 |
|---|---|
| 最高値 | 各属性で最も高い値を持つもの |
| 最低値 | 各属性で最も低い値を持つもの |
| 最低最高以外 | 各属性で最高値，最低値以外の値を持つもの |
| 最高頻度 | 各属性で頻度が最高の値 |
| 最低頻度 | 各属性で頻度が最低の値 |
| 指定属性値 | 属性と属性値を指定したもの |

すなわち，分析の際に有効となる箇所として，平均値などの基準となる値からのズレを重視して,「最高値」「最低値」「最高頻度」「最低頻度」を用意する．また，データ全体の傾向を探るために,「最高最低以外」「指定属性値」を用意する．なお,「指定属性値」は，与えられるデータに特有の属性と属性値のペアを分析者が指定することを想定している．たとえば，総合評価値が 1 から 5 で表されたデータが入力された場合,「総合評価 1」「総合評価 2」などの条件を設定することで，データの傾向を捉えやすくなると考えられる．

### 3.5.2 テキストデータ分析ツールの着目条件

TETDM 上でテキストデータの明示と絞り込みを行うために用意した処理ツール「単語抽出」を図 5 に示す．

テキストデータ分析ツール「単語抽出」による着目条件の設定方法には次の 3 種類がある．

1 分析者による着目条件の設定

2 おすすめによる着目条件

3 単語指定による着目条件

1 の分析者による着目条件の設定方法は，数値データの場合と同様になるが，設定できる項目は,「単語頻


図 5: 単語抽出（マイニング）の操作画面

表 3: テキストデータ分析ツールのおすすめ着目条件

| | |
|---|---|
| 単語頻度 1 | 文章中で頻度が 1 の単語 |
| 単語頻度最大 | 文章中で頻度が最大の単語 |
| 主語頻度 1 | 文章中で主語としての頻度が 1 の単語 |
| 主語頻度最大 | 文章中で主語として頻度が最大の単語 |
| セグメント頻度 1 | セグメント中で頻度が 1 の単語 |
| セグメント共通語 | セグメント頻度が 2 以上の単語 |
| セグメント頻度最大 | セグメント中で頻度が最大の単語 |
| 100 文字以上 | セグメントの文字数が 100 文字以上のデータ |
| 100 文字以下 | セグメントの文字数が 100 文字以下のデータ |

度」「主語頻度」「セグメント頻度」「文字数」となっている．

また，3 の単語指定による着目条件では，入力された単語とその単語を含むセグメントが明示される．

2 のおすすめの着目条件では，数値データの時と同様に，条件設定に悩んだ時に用いられる条件を，表 3 のように用意する．すなわち，平均値などの基準となる値からのズレを重視した条件として，単語を明示する条件として，各頻度が 1 または最大の単語を条件として用意する．また，データの傾向を捉えやすくするための条件として，セグメント頻度が 2 以上の単語を「セグメント共通語」として，セグメントの長さに着目して，文字数が比較的少ないものとして「100 文字以下」，文字数が比較的多いものとして「100 文字以上」の条件を用意する．

## 3.6 着目点の明示機能

3.5 節にて指定された着目条件に合致するテキスト内の箇所をハイライトにより明示する．

すなわち，条件に合致する数値や単語の背景色をオレンジ色で表示し，それらの数値や単語を含むセグメントの背景色を黄色で表示する．

これによって，条件に合致する部分のみに着目することができ，条件に合致するデータについてのみ，より深い分析を行いたい場合には，次節で述べるデータの絞り込みを行う．

表 4: TETDM 上に用意するツールセットの一覧

| ツールセット名 |
|---|
| 1. データの確認と絞り込み |
| 2. テキスト要約 |
| 3. 失礼単語確認 |
| 4. テキスト分類 |
| 5. 単語頻度確認 |

## 3.7 明示データの絞り込み機能

図 4 や図 5 の「データ絞込み」ボタンを押すことで，前節でハイライトされているデータのみにデータを絞り込むことができる．データの絞り込みを行った後は，データ分析の各ツールは，絞り込まれたデータのみの結果を表示する．

そのため，分析者が与えた条件に合致するデータの特徴を見出しやすくなると考えられる．また，この条件の設定と絞り込みを繰り返し行うことで，より詳細な条件に合致するデータについての分析を行うことが可能となる．

## 3.8 着目点の明示と絞り込み機能を利用したツールセット

本節では，3.5 節で述べたテキスト分析ツール「単語抽出（マイニング）」とデータ分析ツール「データマイニング」を，TETDM の既存のツールと組み合わせた図 4 に示す 5 つのツールセット（TETDM 上のツールの組み合わせとなるパネル構成）について述べる．

### 3.8.1 データ確認と絞り込み

ツール「単語抽出」と「データマイニング」のみを利用した最もシンプルなツールセットとして用意する．

### 3.8.2 テキスト要約

ツール「単語抽出」と「データマイニング」に加え，文章要約のツールを利用できるツールセットを用意す

る．キーワードや重要文を確認することで，重要なことが書かれているデータを探す事に役立てられる．

#### 3.8.3 失礼単語確認

ツール「単語抽出」と「データマイニング」に加え，失礼単語を確認するツールを利用できるツールセットを用意する．失礼な表現や否定的な表現を参照することで，改善点の検討が可能になると考えられる．

#### 3.8.4 テキスト分類

ツール「単語抽出」と「データマイニング」に加え，テキストを分類するツールを利用できるツールセットを用意する．データ全体の傾向を眺めることで，どのようなデータが多いかを確認できると考えられる．

#### 3.8.5 単語頻度確認

ツール「単語抽出」と「データマイニング」に加え，単語の情報を確認するツールを利用できるツールセットを用意する．レビューの中でよく使われている単語を確認し，単語の頻度から着目する単語を発見できると考えられる．

## 4 着目点の明示機能の評価実験

本研究で提案する着目点の明示機能が，データ分析の支援に有効かを検証した実験について述べる．

### 4.1 実験内容

楽天市場みんなのレビュー [7] から「ヒーター」のレビューデータ 150 件と,「加湿器」のレビューデータ 100 件を分析してもらい，レビューの総合点数が高くなると予想される新製品の提案をする目的で，商品のレビューデータ[2] を分析してもらった．

実験は，3.5 節で述べた，おすすめの着目条件を利用可能な提案グループと，利用できない比較グループとに分けて行った．被験者は，理系の大学生，大学院生の 10 名で，各グループ 5 名ずつで実験を行った．

[2]本レビューデータは，レビュアの年齢や性別および，商品の評価が 5 段階評価されている数値データとレビューのテキストデータがセットになっている．

表 5: 集められた解釈の数（被験者平均）

| | ヒーター | 加湿器 |
|---|---|---|
| 提案グループ | 10.3 | 10.5 |
| 比較グループ | 7.5 | 8.6 |
| 差 | 2.8 | 1.9 |

#### 4.1.1 実験手順

以下に被験者に提示した，評価実験の手順を示す．

1. 分析に用いるツールセットを選択する．
2. 着目点の明示機能によりハイライトされた数値や単語からデータを絞り込む．
3. 絞り込んだデータの特徴を解釈する．もしくは，さらにデータを絞り込む．
4. 手順 1 から手順 3 を繰り返して解釈をできるだけ多く登録する[3]．
5. TETDM の知識創発機能[4]を用いて，集めた解釈の共通点を見出すことで解釈をまとめる．
6. 共通点が見つからない段階まで解釈をまとめてもらい, それを最終提案とする．
7. 最終的に共通点を絞り込みきれなかった場合は，まとめた解釈を並べて複数文を最終提案とする．

### 4.2 実験結果と考察

#### 4.2.1 「結果と解釈」の登録数

表 5 に，おすすめの着目条件を利用できる提案グループと，利用できない比較グループの被験者が集めた解釈の数を示す．ただし，同一内容の解釈は 1 つとしてカウントしている．

「ヒーター」「加湿器」のいずれのレビューについても，提案グループの方がより多くの解釈を集めることができていた．これは，分析者が自分で思い描く条件だけで分析するよりも，分析者の頭の中にはない条件を提示することで，より多くの解釈を導くことができたためと考えられる．また，分析者が自分で条件を設定するためには，項目の選択，閾値の選択，上限か下限の設定，の 3 つのステップを経る必要があるのに対して，おすすめの条件設定においては，プルダウンメニューの中から 1 つの条件を選択すればよかったため，

[3]TETDM にある「結果と解釈」の登録機能を用いて，データからわかったことを記録してもらう．

[4]登録された結果と解釈を統合して 1 つにまとめるための支援インタフェース

表 6: 最終提案の着目点の数と具体性（被験者平均）

| | 着目点 | 具体性 |
|---|---|---|
| 提案グループ | 3.0 | 2.9 |
| 比較グループ | 2.0 | 1.6 |
| 差 | 1.0 | 1.5 |

表 7: ヒーターの最終提案（提案グループ）

| 回答者 | 最終提案 | 着目点 | 具体性 |
|---|---|---|---|
| 提案 A | 年齢，性別を通して小さく，パワフルであることが利点であり，発送トラブルやデザインの誤解が評価を下げてしまう | 5 | 2 |
| 提案 B | 年齢に関しては，お年寄りは利便性のみに関心が強く，若者は利便性以外にも発送時間に不満がある人が多かった．また，性別のに関しては，女性にとって，デザインとサイズが高評価であった． | 8 | 2 |

表 8: ヒーターの最終提案と評価（比較グループ）

| 回答者 | 最終提案 | 着目点 | 具体性 |
|---|---|---|---|
| 比較 A | この商品はユーザが求める「コンパクトさ」を満たしていると考えられるが，さらにコンパクトさを追求する必要がある | 0 | 1 |
| 比較 B | サイズがコンパクトであり，すぐに部屋が暖かくなること | 1 | 0 |

そのような条件設定の簡易さも，この差につながった可能性があると考えられる．

#### 4.2.2 最終提案の評価の比較

集めた解釈をもとに知識創発を行い，まとめてもらった最終提案についての評価を行った．評価は，データの絞り込みに用いられる属性について，提案の中で言及している属性の数[5] と，具体性として，商品の長所や改善点が具体的に書かれていると判断できる提案[6]を数えた．

結果を表 6 に示す．着目点と具体性のいずれも提案グループの値の方が大きい結果となった．このことから，おすすめの着目条件を用いた被験者の方が，より幅広く具体的な提案につなげることができたことがわかる．

実際の提案の例を表 7 と表 8 に示す．提案グループの被験者の方が，集められている解釈の数が多かったため，それらを用いた幅広い視点からの提案につなげることができたと考えられる．

## 5 結論

本研究では，データにおける平均などの基準値からのズレが大きいデータを，おすすめの着目条件として選択，明示できるようにするとともに，与えた着目点を手がかりとして，データへの着目と絞り込みを繰り返し行える機能を実装した．

評価実験により，おすすめの着目条件が幅広く具体的な考察を行うために有効なことを確認した．

今後は，より効果的なおすすめ条件設定として，複数の属性に関わるおすすめ条件や，データの分布に依存したおすすめ条件の設定を検討していきたいと考えている．

[5] 年齢や性別，評価点数やデザインサイズといった項目に加え，年齢層を示す「若者」や加湿（加熱）能力を示す「パワフル」といった単語も着目点としてカウントしている．

[6] 「インテリアに使える」や「センサーの感度向上」のように，長所や改善点の，方向性や程度がわかる場合にカウントする．

## 参考文献

[1] 赤峯享：ビッグデータ分析でのテキスト情報の活用, 自然言語処理, Vol.20, No.5, p.627, (2013).

[2] 砂山渡, 高間康史, 徳永秀和, 串間宗夫, 西村和則, 松下光範, 北村侑也:統合環境 TETDM を用いた社会実践, 人工知能学会論文誌, Vol.32, No.1, NFC-A, pp.1–12, (2017).

[3] 松尾哉太, 新妻弘崇, 太田学：レビュー解析に基づくユーザ評価の根拠提示の一手法, 情報処理学会研究報告, Vol.35, No.14, pp.1–6, (2014).

[4] 稲垣和人, 吉川大弘, 古橋武：アンケートデータ解析におけるマイノリティの抽出手法に関する検討, 日本知能情報ファジィ学会第 76 回全国大会講演論文集, No.1, pp.383–384, (2014).

[5] 越康治, 高田淑子, 木下英俊, 安藤明伸, 高橋潔, 田幡憲一, 岡正明, 石澤公明:テキストマイニングによる授業評価アンケートの分析:共起ネットワークによる自由記述の可視化の試み, 宮城教育大学情報処理センター研究紀要:COMMUE, No.22, pp.67–74, (2015).

[6] 松本友哉, 砂山渡, 畑中裕司, 小郷原一智：データマイニングとテキストマイニングの連携によるデータ分析支援, 第 15 回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会資料, pp.14–19, (2017).

[7] 楽天みんなのレビュー：
(URL) https://review.rakuten.co.jp/

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-07

# スポット推薦を伴う経路推薦手法の提案

## Proposal of Route Recommendation Method containing Spot Recommendation

柴田 祐樹[1]　高間 康史[1]

Hiroki Shibata[1] and Yasufumi Takama[1]

[1]首都大学東京大学院システムデザイン研究科

[1]Department of System design in Tokyo Metropolitan University

**概要**: 観光案内などでは，訪問すべきスポットだけでなく訪問順序や移動経路も同時に決定する必要がある. この様な，スポット推薦を伴う経路推薦は，観光の分野や日常生活においても需要が高いにも関わらず，問題の定式化の難しさから一般化された手法は十分研究されていない. 目的地を巡回する問題については巡回セールスマン問題(TSP)として多くの研究がなされているが，事前に決定したスポット全てを通る経路を発見するためそのまま適用することはできない. 本研究ではスポット推薦を伴う経路推薦に対し，確率場による定式化を行い，焼きなまし法等を用いた解法を提案する

## 1　研究背景

交通手段, 情報収集手段の発達に伴い, 個人が気軽に未知の土地を訪問できるようになってきている. しかしながら, 個人の趣味嗜好というよりは知名度の高い場所が優先的に選択肢へ上がりやすいという昔からの状況は依然として変わらず, 当該の観光地が混雑する状況を招き, より個人の趣味に一致した場所や経路の選択機会を喪失している. さらにこのような状況が続けば, 観光地としての価値を追求する動きが起こり, その土地本来の良さよりも宣伝性の良さを狙った街作りになる等の弊害も想定され, 訪問者, 現地民双方にとって良い状況とは言えない. このようなことは旅行等の比較的大きなイベントでなくても, 何気ない街や自然の散策, テーマパークの楽しみ方等でも見られる. また, スポットを最短時間でより多く回れば個人の満足度が高くなるというわけでは必ずしもなく, 途中通過する街並みや, 道の景観等も観光, 散策において重要な要素である. 従って, 個人の嗜好や時間的制約などに基づく, 画一的ではない経路推薦手法が必要と考える.

観光経路の推薦では訪れるスポットの選択, 訪れる順番, 総所要時間等に関する要望, スポット以外の道や景観, 混雑状況に対する暗黙的な好みの反映等多様な要求を含んでおり, 一般に取り扱うことが難しく, 定式化の研究は十分になされていない. 目的地を巡回する問題については巡回セールスマン問題 (TSP: Traveling Salesman Problem) [1] として多くの研究がなされているが，事前に決定したスポット全てを通る経路を発見するため，所要時間のバランスを見ながら訪問スポットを選択するといったような調整を行いにくい．ノードに価値を付加し, 総経路負荷が目標値を超えない中で, 価値の総和が最大となる経路を選ぶ問題は Selective Traveling Salesman Problem (STSP) [6] [7]として定式化されているが, 本問題は所要時間が目標値を超過することが許されず, 制限時間の明確でない観光案内へそのまま適用するには不適当である. また, スポットの価値と時間依存性を含む問題を観光経路最適化問題 (ORPS：Optimal Routing Problem for Sightseeing) として定式化することが提案されている[2]. より実用的なシステムとしては CT-Planner [3] が提案されており, 各スポット間を結ぶ最短経路をあらかじめ求めておき, スポットをノード集合とする完全グラフの中で, 遺伝的アルゴリズムを用いて観光経路を求める手法が用いられている.

これらの既存研究は観光経路の決定を, スポットをノードと見立てたグラフ上の経路問題を解くことに帰着している. 文献[2]の定式化はノードの選択や順列の定式化が複雑となっており, 一般性に欠ける. また, 文献[3] の手法はスポットのみに着目しており, 経路に対するユーザの好みに応じた推薦を行うことは想定していない. 本稿では, 経路を構成する辺に対して目的関数を作り, これをエネルギー関数に含む Boltzmann 分布から経路が生成されるとモデル化する. 既存研究ではスポット間の経路は移動負荷を持つものとして幾何学構造と切り離されて考えられているが, 本稿では地図上のすべての道にユーザの好みを反映するという

大規模な問題を想定している. また, 確率場として定式化することで各種統計, 機械学習手法が適用可能となる. 本稿では, 解法として局所最適化法を用い, 確率場による定式化が Simulated Annealing 法の導出を容易にすることを示す. 局所最適化法は滑らかな目的関数を持つ問題の効率的解法として知られており, これを用いることは将来的に有用であると考える. 目的関数の作り方を工夫することで, 局所最適化法を用いた場合でも Local Minimum 問題の発生を緩和できる手法を提案する.

## 2 数学的手法と関連研究について

初めに, 本稿で用いる数学的表記について述べる. また, Boltzmann 分布を用いた定式化を行うので, これについて説明し, その後経路問題を扱った関連研究について述べる.

集合$A$とその直積集合$B_A = \{(x, y)|x \in A, y \in A\}$の関係を$B_A \subseteq A \times A$と表記する. 集合$A$ と集合 $B$の差集合 $C_{AB}$を $C_{AB} = A \setminus B$と表記する. 確率変数は太字立体で書き, その実現値は斜体であらわす. また, 多次元変数$\boldsymbol{x}$に添え字をつけた$x_i \in \boldsymbol{x}$ はその要素を意味する. 多次元離散確率変数 $\mathbf{x} = (\mathbf{x_1}, \mathbf{x_2}, \ldots\ldots \mathbf{x_{|x|}})$の部分空間 $\mathbf{y} \subseteq \mathbf{x}$ に対して, 取りうるすべての状態についての和を$\sum_{\mathbf{y}} f(\mathbf{x})$と書く. また, 特定の, 複数の要素 $\mathbf{x}_i \in \mathbf{x}, \mathbf{x}_j \in \mathbf{x}$については $\sum_{\mathbf{x}_i, \mathbf{x}_j} f(\mathbf{x})$ のように書き, $\mathbf{x}_i, \mathbf{x}_j$の状態が互いに独立でない場合は$\sum_{(\mathbf{x}_i, \mathbf{x}_j)} f(\mathbf{x})$のように書く.

変数 $\mathbf{x}$ についての目的関数$\phi(\mathbf{x})$ を最小化する問題を考える. ここで, (1)式で与えられる Boltzmann 分布を考える.

$$p(\mathbf{x}) = \frac{\exp\left\{-\frac{1}{T}\phi(\mathbf{x})\right\}}{Z_p(T)} \tag{1}$$

$Z_p(T)$ は分配関数であり, (2)式で与えられる.

$$Z_p(T) = \sum_{\mathbf{x}} \exp\left\{-\frac{1}{T}\phi(\mathbf{x})\right\} \tag{2}$$

$\phi(\mathbf{x})$ についての最適解集合を$\mathcal{X}_\mathrm{t}$とする時, (1)式は温度 $T \to 0$ の極限で(3)式となることがわかる.

$$p(\mathbf{x}) = \begin{cases} \dfrac{1}{|\mathcal{X}_\mathrm{t}|} & \text{where} \quad \mathbf{x} \in \mathcal{X}_\mathrm{t} \\ 0 & \text{where} \quad \mathbf{x} \notin \mathcal{X}_\mathrm{t} \end{cases} \tag{3}$$

(1)式の分布を解析的に求め, 直接最適解を得ることは困難であり, 一般に焼きなまし (SA: Simulated Annealing) 法 [5]を用いて近似解が求められる. 焼きなまし法では(4)式を用いて表される条件付き確率を用いて, 温度 T を徐々に低下させながら Markov Chain Monte Carlo 法を適用し, $T \to 0$ での分布を近似する

$$p(\mathbf{y}|\mathbf{x} \setminus \mathbf{y}) = \frac{\exp\left\{-\frac{1}{T}\phi(\mathbf{x})\right\}}{\sum_{\mathbf{y}} \exp\left\{-\frac{1}{T}\phi(\mathbf{x})\right\}} \tag{4}$$

次に, 観光案内に関連する経路問題について述べるが, これらを Boltzmann 分布により定式化, 解く方法は 3 節の提案手法で述べる.

与えられたノードをすべて巡回する最短経路を求める問題は巡回セールスマン問題 (TSP) として広く知られている. TSP である必要条件は各ノード間の負荷が定義されていて, すべてのノードを巡回する経路のみを解として認めることである. 2-opt 法 (k-opt 法) [4] は TSP を解く局所最適化法の 1 つであり, Fig. 3 に本手法が動作する様子を示す. 2-opt 法は, 隣接しない 2 つの辺を選び, 経路が開始点から終止点まで連続している状態を保つ入れ替え方のうち, 全体の負荷をより小さくするものへ置き換える操作を繰り返すことで最適化を行う. しかし, 観光案内では, すべてのノードを巡回する経路が求められるわけではないので, TSP として定式化することはできない.

1 節で述べた文献[2]では STSP の発展型として, 観光経路最適化問題を ORPS として定義している. また, ORPS が NP 困難であることを示し, 解析的手法による厳密解法と発見的手法による近似解法についても述べている. しかしながら, 定式化はノード集合を基準としたものであり, 本稿が提案する辺集合による目的関数の定式化に比べ条件設定が必要であり煩雑となっている. CT-Planner[3] では, 最適化手法よりもユーザとのインタラクティブ性を重視している. ユーザの好みを知らない状態から開始し, ごく少数の質問に対する回答に基づき推薦をしながら, 段階的にユーザが経路を最適化できるような設計指針を持っている. また, ユーザが観光地情報を拡張可能なインタフェースも提供しており, 所定の書式に従った表形式ファイルを用意することで新たなスポットを追加できる. 最適化手法としては遺伝的アルゴリズムを用いており, データベース内にある任意の観光スポット間の移動負荷をあらかじめ求めて置き, 滞在時間と, ユーザごとに異なる価値が与えられた各観光スポットをノード, ノード間の辺に重みとして移動時間が割り当てられた完全グラフとして目的関数を定義, ユーザが指定した所要時間以下で価値の総和が最大となる経路を求めている.

# 3 提案手法

本稿で対象とする問題は, 総移動時間及び目標総移動負荷の制約がある中で, ユーザの経路に対する満足度を最大化するものを推薦することである. 推薦スポットは, 従来ノードに価値を付加することで表現していた代わりに, 本稿では仮想的な辺, あるいはスポット内の実在経路上, 例えば博物館の中の通路等に価値を割り当てることで表現する. Fig. 1 にスポット内へ仮想経路を配置した図を示す. スポット内の仮想経路の価値はユーザの好みに応じて変化させることで, 喫茶店のような滞在時間の変動が大きいスポットにも冗長な経路を巡回する問題として対応できる. 3.1 節では辺に対して定義されるこれらの重みから, 目的関数を定義し, これをエネルギー関数とした Boltzmann 分布を用いて, 経路は確率的に生成されるとモデル化する. 3.2 節では TSP に適用した場合, 3.3 節では局所最適化法を適用する場合の目的関数を紹介し, 3.4 節で 3.3 節までのものに加え, 辺に負荷された価値と総所要時間の制約を考慮した目的関数を示す.

## 3.1 Boltzmann 分布と経路の定義

地図を構成するノード集合を $\mathcal{N}$ とする. 各ノードを結ぶ辺から構成される, 終始点を通る有効な経路を含むベクトルを, $\boldsymbol{e} = (e_1, e_2, \dots e_i \dots e_{|\boldsymbol{e}|}), e_i \in \mathcal{N} \times \mathcal{N}$ と表す. Fig. 2 に示す例では, $\mathcal{N} = \{1,2,3,4,5,6\}$, $|\boldsymbol{e}| = 5$ の場合に以下のようになる.

$$\boldsymbol{e} = ((1,2),(2,3),(3,4),(4,5),(5,6))$$

3.3 節で示す経路への辺の追加などに対応するため, 辺ベクトルは経路に含まれない辺を含む. 例えば, $|\boldsymbol{e}| = n$の場合以下のようになる.

$$\boldsymbol{e} = \left((1,2),(2,3),(3,4),(4,5),(5,6), e_6, e_7 \dots, e_n\right)$$

$e_1$から$e_5$のみで有効な経路を構成する. 辺の方向は第 1 ノードから, 第 2 ノードへ向かうものとする. $\boldsymbol{e}$の確率変数を$\mathbf{e}$とし, その分布$p(\mathbf{e})$を(5)式で与える.

$$p(\mathbf{e}) = \frac{\exp\left\{-\frac{1}{T}\phi(\mathbf{e})\right\}}{Z_p} \tag{5}$$

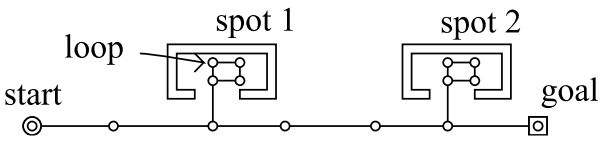



Fig. 1. Virtual redundant loop in spots

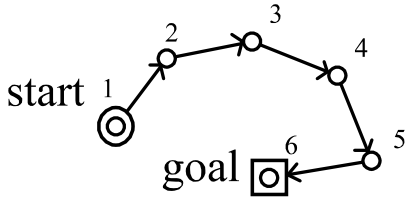



Fig. 2 An example of route.

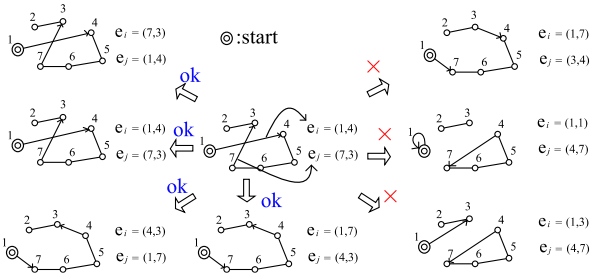



Fig. 3 The restriction to the edge's status.

$\phi(\mathbf{e})$ は目的関数であり, 経路 $\mathbf{e}$ に対する制約を定義する. $\phi(\mathbf{e})$の値が小さいほど$p(\mathbf{e})$の値が高まり, そのような経路が選択される確率が高くなることがわかる.

## 3.2 TSP への適用

TSP において目的関数は経路上の辺$\mathbf{e}_i \in \mathbf{e}$に割り当てられた重み$f_c(\mathbf{e}_i)$の総和で与えられ, (6)式となる. 一般的にはグラフ上の辺に割り当てる値のことを重みと呼ぶが, 3.4 節で説明するように, 本稿ではユーザの好みも辺に割り当てることから重みは多次元量となり, それぞれ負荷, 好みのように呼ぶことにする.

$$\phi(\mathbf{e}) = \sum_{\mathbf{e}_i \in \mathbf{e}} f_c(\mathbf{e}_i) \tag{6}$$

TSP を解くために提案された 2-opt 法を用いた SA 法を導出する. 局所最適化対象の 2 辺を確率変数$\mathbf{e}_i \in \mathbf{e}, \mathbf{e}_j \in \mathbf{e}$で表し, これらを除いた辺ベクトルの部分空間 $\mathbf{e} \setminus \{\mathbf{e}_i, \mathbf{e}_j\}$ により条件つけられた, (5)式に対する条件付確率は(7)式により与えられる.

$$p(\mathbf{e}_i, \mathbf{e}_j | \mathbf{e} \setminus \{\mathbf{e}_i, \mathbf{e}_j\}) = \frac{\exp\left\{-\frac{1}{T}\phi(\mathbf{e})\right\}}{\sum_{(\mathbf{e}_i, \mathbf{e}_j)} \exp\left\{-\frac{1}{T}\phi(\mathbf{e})\right\}} \tag{7}$$

ここで, $\mathbf{e}_i, \mathbf{e}_j$の取りうる状態について考える. 経路が開始点から終止点まで連続している状態を保つ必要があり, その遷移状態の組み合わせは Fig. 3 に示す “ok” と書かれた 4 つしかない. 許されない遷移状態は, 開始点を含まない閉じた経路ができてしまうもの, 経路の巡回方向と一致しない辺の

方向となるものであり, その一部のみを例示した. ある辺 $\mathbf{e}_t \in \mathbf{e}$を構成する 2 つのノードを$a_t, b_t$と表し, (8)式に $\mathbf{e}_i, \mathbf{e}_j$ の取りうる状態を示す.

$$(\mathbf{e}_i, \mathbf{e}_j) \in \begin{Bmatrix} \big((a_i, b_i), (a_j, b_j)\big), \big((a_j, b_j), (a_i, b_i)\big), \\ \big((b_i, b_j), (a_i, a_j)\big), \big((a_i, a_j), (b_i, b_j)\big) \end{Bmatrix} \quad (8)$$

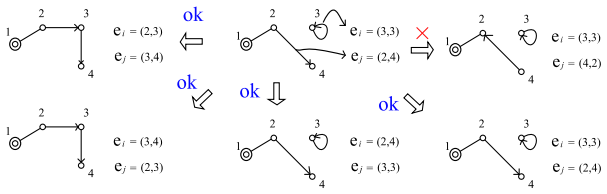
$\mathbf{e}_i, \mathbf{e}_j$に対して変更があった場合, 経路上のほかの辺に対しても巡回方向が正しくなるように修正を加える.

温度$T$ を減少させながら, あらゆる辺の組み合わせについて(7), (8)式で定義される確率にしたがい辺を入れ替えることを繰り返すことで, 最適解を近似的に得ることができる.

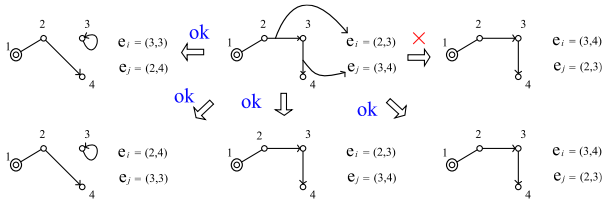
## 3.3 辺の経路への追加と除外

本稿では局所最適化法により観光経路問題を扱う方法を提案する. 局所最適化法では一度に扱える変更幅が小さく, 複雑な経路問題に対しては一般的に適用することができない. 本稿では辺の, 経路への追加および除外過程を導入することでこの課題に対処する. 経路を局所最適化法により少しずつ拡張していくためには, 途中, 地図上に実在しない非存在辺を一時的に追加する必要がある. しかし, 計算終了時, つまり$T = 0$の時には非存在辺がすべて実在辺に置き換わっている必要があるため, 非存在辺における罰則$f_p(e)$を導入し, 負荷$f_c(e)$と合わせ目的関数$\phi(\boldsymbol{e})$を(9)式とする. また, 非存在辺においては$f_c(e) = 0$ とする.

$$\phi(\boldsymbol{e}) = \sum_{e \in \boldsymbol{e}} f_c(e) + \sum_{e \in \mathbf{e}} f_p(e) \quad (9)$$




(a) Adding




(b) Removing

Fig. 4 The restriction of edges' status to add or remove an edge from the route.

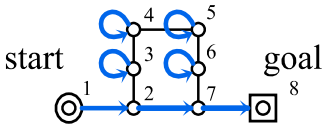



Fig. 5 An example of self loop edges.

罰則$f_p(e)$に関して, $\boldsymbol{s}$ を非存在辺$e$の始点, 終点を結ぶ実在最短経路, これに含まれる実在辺を$s_j \in \boldsymbol{s}$ とし, (10)式を満たすようにすることで, $T = 0$ となったときに非存在辺はすべて実在辺に入れ替わる. ただし, 実在最短経路が存在しない場合, $f_p(e) = \infty$ とする.

$$f_p(e) > \begin{cases} \sum_{s_j \in \boldsymbol{s}} f_c(s_j) & where \quad |\boldsymbol{s}| \neq 0 \\ \infty & where \quad |\boldsymbol{s}| = 0 \end{cases} \quad (10)$$

例えば, Fig. 2 のにおいて, $e = (1,4)$ の時, 対応する実在最短経路は. $\boldsymbol{s} = \big((1,2), (2,3), (3,4)\big)$ となる.

(10)式を満たす具体的な例として(11)式があり, 本稿ではこれを用いた. $C_{pb}, C_{pc}$ はともに定数である.

$$f_p(e) = C_{pb}\left(\sum_{s_j \in \boldsymbol{s}} f_c(s_j)\right) + C_{pc} \quad (11)$$

また, 本稿では 2-opt 法により辺の追加と除外を考慮するが, この場合, 辺の取りうる状態として Fig. 3 に加え, Fig. 4 も許可される. (a)は経路に含まれないノード 3 上に自己ループとなる辺があらかじめ配置されている下で, この辺と経路に含まれる辺について状態の変化を考え, 経路にとって新しいノード 3 に対して辺が 1 つ追加される過程を表現している. 結果として経路に含まれないままの状態も許可される. また, (b) は経路内の隣接する 2 つの辺の状態を, 一方の辺が自己ループとなるように変化させることで, 1 つのノードが経路から外れる除外過程を表現する. 3.2 節と同様にここでも許可される遷移状態を “ok” と表し, 許可されないものについては一部の例を示した. この(a), (b)の方法は, 初期状態として経路に含まれないノードに自己ループ辺を与えておけば, 経路の拡張縮小を繰り返してもノード上に自己ループ辺がある状態を保つことができる. 例えば, Fig. 5 は地図上に青で辺ベクトルに含まれる要素を示している, 自己ループ辺がノードにそれぞれ 1 つずつ配置されていた場合の例で, この図において以下のようになる.

$$\boldsymbol{e} = \big((1,2), (2,7), (7,8), (3,3), (4,4), (5,5), (6,6)\big)$$

各ノードを複数回通る経路を生成可能とする場合には, この自己ループ辺はノードごとに複数個用意しておく.

Fig. 4 (a), (b)に描かれている状態の変化は, 同じ規則に従っており, $\mathbf{e}_i, \mathbf{e}_j$の取りうる状態を整理すると, (12)式の様になる.

$$(\mathbf{e}_i, \mathbf{e}_j) \in \left\{ \begin{matrix} \big((a_i, b_i), (a_j, b_j)\big), \big((a_j, b_j), (a_i, b_i)\big), \\ \big((a_j, b_i), (a_i, b_j)\big), \big((a_i, b_j), (a_j, b_i)\big) \end{matrix} \right\} \quad (12)$$

(12)式を見れば, $a_j = b_j$もしくは$a_i = b_i$の場合に自己ループ辺となり, 辺の追加による経路の拡張に対応する. そうでない場合, $\mathbf{e}_i, \mathbf{e}_j$ は隣接するので, $a_j = b_i$もしくは, $a_i = b_j$となる場合は辺の削除に対応し, 経路の縮小を考慮できていることがわかる. 自己ループ辺を配置しておくことで, 経路に対して辺が追加, 除外されることはあっても, $\mathbf{e}$ を構成する要素数は固定されることから, 辺ベクトルとして扱うことができ, 経路を含む辺の集合に対する確率場を, 経路の拡張縮小を考慮したうえでも定式化することができる. なお, 自己ループ辺を配置しておくことは, これら定式化における利便性のために導入したもので, 実装上必ずしも必要なものではない.

なお, (10)式を満たさない場合, 非存在辺が最終的に残る可能性がある. $f_p(e)$を幾何学距離の関数で近似することも考えられるが, (10)式を常に満たすために, $f_p(e)$を過大に設定する必要がある. この時, 非存在辺への遷移確率を低下させ, 経路探索を行いづらくなる. すなわち, $f_p(e)$ が(10)式の右辺に近いほど経路探索を行いやすくなる. $f_p(e)$はすべてのノード対に対して計算しておく必要があり, その記憶量, 計算量は$O(|\mathcal{N}|^2)$となる. しかしながら, 通常計算開始時に$f_p(e)$は一度だけ計算しておけばよいため, この計算負荷は大きな問題とならない.

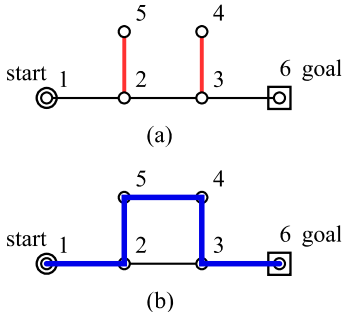



Fig. 6 An example where we have to add $f_d(s_j)$

## 3.4 緩和 STSP

本項ではユーザの辺に対する好みを$f_d(e)$ で与える. 既存研究ではノードに対してスポットを表現する値が付加されていたが, 本研究では経路負荷を辺に対する関数$f_c(e)$として与えており, これとの一貫性を取るために辺に対してスポットを割り当てる. 同一ノードを複製した 2 ノード間に距離が 0 の辺を考えることで, ノードに対して割り当てたものと等価な問題を考えることができる. ユーザの好み$f_d(e)$を取り入れ, 目標移動負荷値$C_c$を導入したエネルギー関数を(13)式で与える.

$$\phi(\mathbf{e}) = \frac{\beta}{2\sigma^2}\{F_c(\mathbf{e})\}^2 + (1-\beta)\left\{F_c(\mathbf{e}) - \frac{\sigma^2}{2}\right\} - \sum_{\mathbf{e}_i \in \mathbf{e}} f_d(\mathbf{e}_i) + \sum_{\mathbf{e}_i \in \mathbf{e}} f_p(\mathbf{e}_i) \quad (13)$$

ただし, 条件変数 $\beta$, 総移動負荷$F_c(\mathbf{e})$は(14)式で与えられる.

$$F_c(\mathbf{e}) = \left| C_c - \sum_{\mathbf{e}_i \in \mathbf{e}} f_c(\mathbf{e}_i) \right|, \quad \beta = \begin{cases} 1 & where\ \sigma^2 > F_c(\mathbf{e}) \\ 0 & where\ \sigma^2 \le F_c(\mathbf{e}) \end{cases} \quad (14)$$

STSP では$\phi(\mathbf{e})$が$C_c$ で表される目標移動時間を少しでも上回ることを認めないが, 観光における経路推薦では妥当な近傍解は許容されることを考え, (13)式のように総移動負荷$F_c(\mathbf{e})$に対して最小値を持つエネルギー関数を導入する. また, (13)式は移動負荷の増加に見合う $f_d(\mathbf{e}_i)$の値が得られるスポットに対し経路の拡張を許可する. $\sigma^2$は拡張規模の許容範囲を表現する. また, $f_d(\mathbf{e}_i)$ を導入したことにより, $f_p(e)$の定義式(10)式を(15)式に変更する. また, (10)式を満たすための近似式として(11)式の代わりに(16)式を用いる.

$$f_p(e) > \sum_{s_j \in \mathbf{s}} \{f_c(s_j) + f_d(s_j)\} \quad (15)$$

$$f_p(e) = C_{pb}\left(\sum_{s_j \in \mathbf{s}} \{f_c(s_j) + f_d(s_j)\}\right) + C_{pc} \quad (16)$$

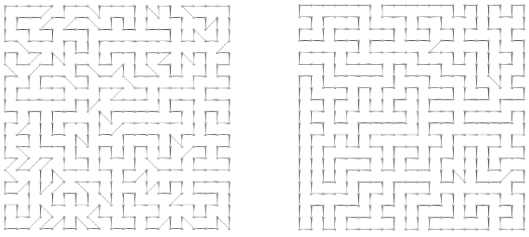
$f_d(s_j)$ は目的関数に対して減算されているので, 罰則関数に加算することは奇妙に思える. Fig. 6 に $f_d(s_j)$ を加算しなければならない状況の例を示す. (a)において, 赤く示されている辺は高い価値が設定されている辺であり, $f_d((2,5)) = f_d((5,2)) = f_d((3,4)) = f_d((4,3)) = 3$ ほかの辺$e$については

$f_d(e) = 0$とする. 実在辺の負荷は 1 とする. (13)式において, $\sigma^2 \to 0, C_c = 0$, (付録に詳細を記載) また(11)式に$C_{pb} = 1, C_{pc} = 1$を用いるとして, Fig. 6 (b)に青で描かれた, 非存在辺を含む経路$\mathbf{e}^{(*)} = \big((1,2),(2,5),(5,4),(4,3),(3,6)\big)$の(13)式の値は$\phi(\mathbf{e}^{(*)}) = 2$, 望まれる実在辺のみの経路$\mathbf{e}^{(t)} = \big((1,2),(2,3),(3,6)\big)$に対する(13)式の値は$\phi(\mathbf{e}^{(t)}) = 3$となり, $\phi(\mathbf{e}^{(*)}) < \phi(\mathbf{e}^{(t)})$であることから$\mathbf{e}^{(*)}$が選ばれてしまうことがわかる. 同じ条件で(16)式を用いた場合は, $\phi(\mathbf{e}^{(*)}) = 8, \phi(\mathbf{e}^{(t)}) = 3$ であり, $(\mathbf{e}^{(*)}) > \phi(\mathbf{e}^{(t)})$ となるので, $\mathbf{e}^{(t)}$ が選ばれることがわかる. また, (11)式の罰則定数$C_{pb} = 1$, $C_{pc} = 1$をいくら大きくしても, $f_d(e)$の大きな場所があれば, 非存在辺が残る可能性をなくすことはできない. 定性的には, 非存在辺を通ってでも, 好みの経路を通った方が目的関数の値を小さくできるような状況を避ける必要がある, ということが言える

## 4 評価実験

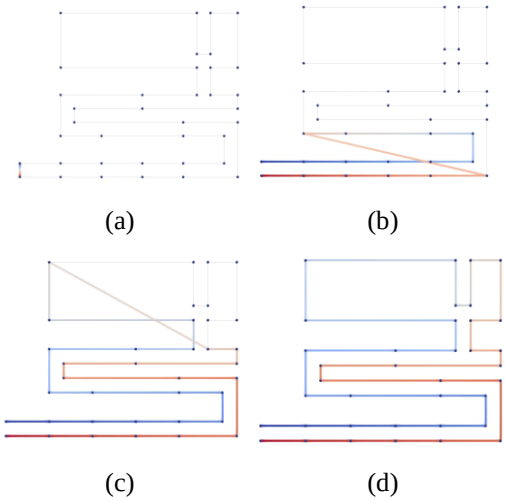
3.2 節で説明した TSP へ適用した結果を Fig. 7 に示す. (a)は$T = 0$で計算した結果で, 2-opt 法に一致するものであり, (b)は$T = 0.093$ で計算した結果で焼きなまし法となる. 20 行 20 列の幅 1 で置かれた等間隔配置ノードからなる, ノード数 400 の完全グラフに対し, 全ノードを通り負荷最小の経路を求めている. この時, すべてのノード対に幾何学距離に等しい負荷が設定されており, 斜めの辺が 1 つもない経路が最適解となる. (b)の焼きなまし法を用いた結果がより最適解に近いことがわかる.

Fig. 8 に辺の追加と除外の仕組みが機能している様子を示す. 実在辺は薄い灰色で表示されており, これら以外のノード対を結ぶ辺はすべて非存在辺である. 開始点と終止点が設定されており, 開始点と終止点がつながる経路のうち目的関数(9)式が最小となる経路を求める. この問題の最適解は, 37 ノードすべてを実在辺のみで通る経路である. (a)に初期状態を示しており, 初期経路として開始



(a) (b)

Fig. 7 The result of TSP with SA and Markov Chain Monte Carlo.



(a) (b)

(c) (d)

Fig. 8 The result of edge adding and removing. $T = 0$

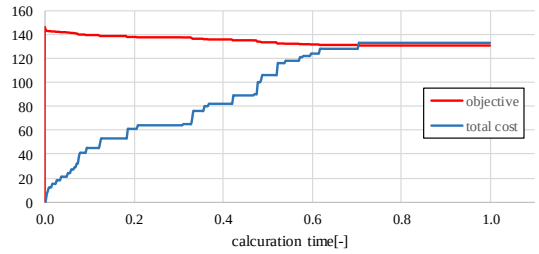



Fig. 9 The graph of total value of $f_c(e)$ in the route and objective function. Red is the objective function while blue denotes the total value of $f_c(e)$.

点と終止点を直接結ぶ非存在辺からなる経路を与えている. $f_p(e)$ を定義する(11)式のパラメータとして, $C_{pb} = 1.1, C_{pc} = 0.1$ を用い, $T = 0$ とした.

この問題は, 罰則関数$f_p(e)$を幾何学距離で近似した場合にうまくいかない複雑な経路上において, 計算終了時に非存在辺が残らないことを検証することを目的としたものである.

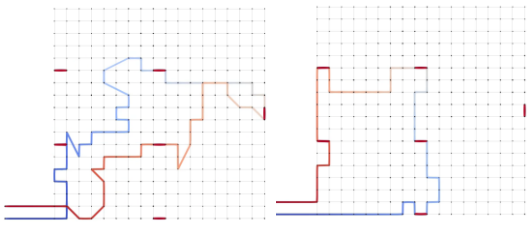
Fig. 8 の(b), (c), (d) はそれぞれ 40, 200, 400 ステップにおける経路である. 開始点に近い辺を赤, 終止点に近い点を青で, グラデーションを用いて可視化している. 局所最適化を繰り返すうちに非存在辺と実在辺の追加と除外を繰り返し, 最終的に実在辺のみが残っていることがわかる. また, Fig. 9 にエネルギー関数と総コストの変化過程を示す. 総コストの上昇に対して, 非存在辺が消滅することでエネルギー関数は常に減少していることがわかる.

Fig. 10 に(13)式のエネルギー関数で表される緩和 STSP を解いた結果を示す. ここでは問題を簡潔にするため, 18 行 18 列の等間隔におかれたノードからなるマス目上の地図を想定した. ノード数は

終始点を含め 326 である. $\sigma = 5, C_c = 60$ とし, 赤く表示されている部分は$f_d(e) = 2, f_c(e) = 1$ の辺であり, そのほかの薄い灰色で描画された辺は $f_d(e) = 0, f_c(e) = 1$ の辺である. これら以外のノード間をつなぐ辺はすべて非存在辺である. (a)は計算開始直後の状態であり, 温度$T$ が高いことにより発見的な経路探索を行っていて, 非存在辺も含んでいる. 斜めの辺はすべて非存在辺である. (b)は計算終了時の状態であり, 非存在辺はなくなっており, 負荷の目標値として妥当な経路でスポットを巡回する経路を生成していることがわかる.

## 5 結論

観光経路推薦に要求される, スポットの選択性, 目標移動負荷および時間, 経路の推薦を含む問題に対して, 辺の集合に対する確率場を Boltzmann 分布を用いてモデル化し, 局所最適化法により経路を生成する方法を示した. また, 本手法は, 辺に対するユーザの好みに応じた経路生成が可能なこと, ノードであらわされるスポットに対しても仮想的な辺を配置することで経路推薦が行えることも示した. 局所最適化法により効率的に解の探索が行える本提案手法は, スケーラビリティに優れるため, 広域を対象とした観光案内や, 経路上のユーザの好みまで反映させる大規模な問題にも適用可能と考える. 今後の展望としては棄却, 重点サンプリング[8]を導入し, 近傍経路に対する確率場の評価を重点的に行い, 遠方への辺検索に要する無駄な探索を抑制することを検討している. また, 現在 2 辺の組み合わせによる最適化だけを考慮しているが, より多くの組み合わせを用いた場合について検証する. さらに, スポットの価値が時間依存性を持つ場合にも対応できるようにする. 最適化手法だけではなく, 実際の経路推薦に適用された場合のユーザの満足度等, ユーザとのインタラクティブ性を考慮し, 経路推薦手法としての精度評価指標も検討する予定である.



(a) (b)

Fig. 10 The result of reducted STSP. The red edges have the weight of $f_d(e) = 2, f_c(e) = 1$ while graies have $f_d(e) = 2, f_c(e) = 1$ of it.

## 参考文献

[1] G. Gutin, A. Punnen, “The traveling salesman problem and its variations,” Springer Science & Business Media, 2006.

[2] 松田善臣, 名嘉村 盛和, 姜 東植, 宮城 隼夫, 最適観光経路問題とその解法, 電気学会論文誌 C (電子・情報・システム部門誌), 2004, Vol. 124, No. 7: pp. 1507-1514.

[3] 倉田陽平, 有馬貴之, 対話的旅行計画作成支援システムの実装と評価, 第 25 回日本観光研究学会全国大会, 日本観光研究学会全国大会学術論文集, 2010, pp. 173-176.

[4] K. Helsgaun, “General k-opt submoves for the Lin–Kernighan TSP heuristic,” Mathematical Programming Computation, 2009, Vol. 1, No. 2-3: pp. 119-163.

[5] S. Kirkpartrick, C. D. Gelatte, Jr., M. P Vecchi, “Optimization by simulated annealing,” Science, 1983, Vol. 220, No. 4598: pp. 671-680.

[6] G. Laporte, S, Martello, “The selective travelling salesman problem,” Discrete applied mathematics, 1990, Vol. 26, No. 2-3: pp. 193-207.

[7] D. Feillet, P. Dejax, M. Gendreau, “Traveling salesman problems with profits,” Transportation science, 2005, Vol. 39, No. 2: pp. 188-205.

[8] C. M. Bishop, “Pattern Recognition and Machine Learning,” Springer, 2006.

[9] 倉田陽平, 原辰徳, インターネット上での対話的旅行プラン作成支援サービス とその展開可能性, サービス学会第 2 回国内大会, 2014, pp. 191-194.

## 付録

(14)式の条件のもと, $\sigma^2 \to 0$ の極限を(13)式に適用すると, (17)式となる.

$$\lim_{\sigma^2 \to 0} \phi(\mathbf{e}) = F_c(\mathbf{e}) - \sum_{\mathbf{e}_i \in \mathbf{e}} f_d(\mathbf{e}_i) + \sum_{\mathbf{e}_i \in \mathbf{e}} f_p(\mathbf{e}_i) \quad (17)$$

また, $F_c(\mathbf{e})$ 中の定数 $C_c$ を 0 とすれば, (18)式となる.

$$\phi(\mathbf{e}) = \sum_{\mathbf{e}_i \in \mathbf{e}} f_c(\mathbf{e}_i) - \sum_{\mathbf{e}_i \in \mathbf{e}} f_d(\mathbf{e}_i) + \sum_{\mathbf{e}_i \in \mathbf{e}} f_p(\mathbf{e}_i) \quad (18)$$

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-08

# 多様な観点から相手の考慮を促すプレゼント選択支援手法の提案

# Supporting Gift Selection to Encourage Gift-receiver's Consideration from Various Perspectives

西野 沙紀 [1]　　松下 光範 [1*]
Saki Nishino[1]　　Mitsunori Matsushita[1]

[1] 関西大学 総合情報学部
[1] Faculty of Informatics, Kansai University

**Abstract:** The purpose of this research is to encourage consideration of a gift-receiver from various perspectives in the process of gift selection. People often purchase items as gifts for others such as a partner and friends. To select a gift for them, e-commerse sites that recommend gift items are available, however, it is difficult to select an appropriate one because these sites recommend items without considering hobbies and preferences of the gift-receivers. To solve the problem, we propose a system that encourages consideration about the gift-receiver. To achieve this, the proposed system provides questions about the gift-receiver and facilitates the deeper think of gift selection. With the proposed system, we conducted an experiment to observe the gift selection process of the participants. As the result, we confirmed that the participants' utterances about gift-receivers are increased.

## 1　はじめに

インターネットの普及に伴い，Amazon や楽天市場などのショッピングサイトが数多く登場し，日常的に利用されるようになってきている．ユーザはそれらのショッピングサイトを利用することで多様な商品を選択できるようになったものの，商品選択の幅が広がったことにより，それらの中からユーザが求める商品を見つけることが容易ではなくなっている．この問題を解決するために，ユーザの嗜好に合った商品推薦を行う技術が提案されている．例えば，過去の購買履歴データからユーザごとの好みを推測し，ユーザの嗜好に合った商品を推薦する手法が提案されている [1]．また，Amazon では欲しい商品を選択した際に，「この商品を買った人はこんな商品も買っています」といった商品推薦が行われている [2].

個人の購買行動に対する支援は，これらの技術により支援されているが，商品の購買行動は個人のためだけではなく，他者へプレゼントを贈る際にも行われる．こうした購買の支援として，プレゼントに適した商品を集めた COCOMO（https://cocomo.to/）やギフト・ディノス（https://www.dinos.co.jp/gift/），おくりものナビ（https://gift.rakuten.co.jp/）などのサイトがサービスを行っている．これらのサイトでは，プレゼントを探す際に，贈る相手の年齢，イベントや場面を選択することができ，それぞれにあった商品が推薦される．また，プレゼントに関するまとめサイトやアンケートに基づくサイトなども多く存在し，「20 代おすすめのプレゼント」や「もらって嬉しい誕生日プレゼントランキング」などを提供している．

しかし，これらのプレゼントサイトによって推薦された商品は，プレゼントを贈る相手の趣味や嗜好などについて考慮されていないため，必ずしも贈る相手に適した商品であるとはいえない．そのため，これらのサイトを利用しても，プレゼントを贈る側（以下，贈り手と記す）が納得した商品を選択できるとは限らない．

このような背景の下，本研究では，プレゼント選択において，相手について「考える」行為に着目してシステムを設計する．提案システムは，従来のプレゼント推薦システムとは異なり，贈り手がプレゼントを選択する際に，プレゼントを贈る相手について多様な観点から考えさせることで，贈り手自身がプレゼント選定の過程に納得することを企図したプレゼント選択の支援を狙う．

*連絡先：関西大学総合情報学部
〒 569-1095 大阪府高槻市霊山寺町 2-1-1
E-mail: mat@res.kutc.kansai-u.ac.jp

## 2 関連研究

水野らは，検索エンジンにユーザが入力した一般的な単語に対して，クエリの拡張を行うことで，ユーザの嗜好に合わせた商品を検索・推薦が可能な手法の提案を行っている [7]．嗜好情報として利用したユーザのブログを用いたクエリ拡張だけでなく，ドメイン内の専門単語を利用した拡張を行った．また，ユーザに対してプロファイルを作り，商品検索を行ったところ，システムを使用することで，発見が困難である商品を発見することが可能になった．しかし，ユーザプロファイルに基本クエリが含まれていない場合は，拡張クエリの生成が不可能であることが課題として挙げられている．村上らは，ユーザの嗜好を反映するために，テキストマイニングと属性生成手法である Category-guided Adaptive Modeling 法（以下，CAM 法と記す）を組み合わせた商品推薦システムを提案した [6]．CAM 法を用いて個人感性モデルのひとつである嗜好モデルを商品の選好情報から構築している．商品に対して「可愛い」といった感性語と嗜好モデルに基づいて，ユーザの嗜好に適した商品を推薦している．コスメを対象に商品推薦システムを構築し，利用者に実験を行ったところ，推薦した商品に対し，「半分以上が購入したい」，「すでに購入している」といった結果が得られた．しかし，これらは自身のための商品購買の支援であり，他者へのプレゼント選択は考慮されていない．

田口らは，様々なオンラインショッピングサイト（以下，EC サイトと記す）の商品レビューを用いて，プレゼントに適した商品を推薦するシステムを提案した [5]．商品がプレゼントに適しているかを判定するために，プレゼントとして扱われた商品のレビューから TF-IDF 特徴量を抽出し，Support Vector Machine （以下，SVM と記す）のモデルを作成した．このモデルを用いたシステムでは，ある商品カテゴリ（e.g., インテリア）を検索した際に，その中の商品を SVM を用いてプレゼントに適しているかを判定し，判定された商品を提示する．このシステムでは，プレゼントに適した商品の選定が行われたものの，プレゼントを贈る対象者やプレゼントを贈る場面については考慮されていない．盆子原らは，相手の趣味や嗜好を考慮し，贈り手が納得したプレゼント選択をするために，EC サイトの商品ジャンルにある，ツリー構造に着目したシステムを提案した [4]．EC サイトでは上位ジャンル（e.g., ファッション）から下位ジャンル（e.g., 服，時計）へとジャンルを辿りながら商品を選択するのが一般的であるが，このシステムでは下位ジャンルから上位ジャンルへの商品探索を可能とし，ユーザが多様な商品ジャンルに気づくことができる工夫を施している．提案システムを用いて実験を行ったところ，探索回数と探索時間が EC サイト型システムよりも増加した．様々なジャンルを選択することで，相手の趣味や嗜好を考慮した探索が行うことが可能になった．空中らは，相手に贈る商品の心象が定まっていない状態の贈り手に対し、無意識下の要求に気づかせることで、きっかけなしでは見つけることの難しい商品の発見を可能にするための支援を行った [3]．贈る相手の心象と，商品ジャンルを結びつけて提示することで，贈り手自身の内省を促し，プレゼントの探索過程における充実感を高めるシステムを提案している．システム上で相手に関するキーワードと商品情報を結びつけ，このキーワードにあった商品ジャンルの円が拡大される．これにより，新たな商品ジャンルに対する気づきを得ることができる．

本研究では，これらの研究を参考に，自身のための商品購買ではなく，他者へのプレゼントとしての商品購買に対する支援を行う．その中でも，プレゼントを選択する過程に着目し，贈る相手についての考慮を促すことで，贈り手が納得した商品を選択することを目指す．本稿では，従来のような商品の推薦を行うシステムではなく，贈り手に対して贈る相手の考慮を促すためのきっかけを与えるシステムの実現を目指す．

## 3 提案手法

### 3.1 相手について考慮させる手法

相手を考慮するきっかけがない状態で贈り手が多様な観点から相手について考慮することは難しい．そこで，本稿ではきっかけを与える手段として，システムが相手についての質問を提示する方式を用いる．贈り手が質問をきっかけに，相手について考慮したプレゼントを選択できることを目指す．

相手について考えさせるための質問項目を定めることを目的とした予備実験を行った．大学生 2 名を実験協力者とし，特定のプレゼントを贈る相手を想定させた上でプレゼントを選択してもらった．実験は商品検索の際に思っていること，考えていることを可能な限り発話してもらう発話思考法を用いて，選択過程を観察した．相手がよく身に着けているものや，相手の持ち物で傷んでいるものに関する発話が得られたことから，相手について考慮した様子が観察された．それらの発話をもとに，相手についての質問を 18 項目決定した．それらの質問項目を表 1 に示す．

### 3.2 提案システムの全体像

質問をきっかけとし，贈り手自身がキーワードを生み出すことで商品検索を行うシステムを提案する．以下に想定されるインタラクションの例を示す．

表 1: 相手についての質問項目

| 番号 | 質問 |
|---|---|
| 1 | 相手の環境（大学院生，バイト先，一人暮らしなど） |
| 2 | 相手に起こったイベント（就職，進学，結婚，留学など） |
| 3 | 相手の趣味 |
| 4 | 相手がよく身に着けているもの |
| 5 | 相手の持ち物で傷んでいるもの・買い替え時のもの |
| 6 | 相手の好きなブランド |
| 7 | 相手の好きなキャラクター |
| 8 | 相手の好きな色 |
| 9 | 相手の好きな芸能人 |
| 10 | 相手が欲しいもの |
| 11 | 相手の好きな食べ物 |
| 12 | 相手が大切にしているもの |
| 13 | 相手が好きなアーティスト・曲 |
| 14 | 相手が好きなテレビ |
| 15 | 相手が憧れている有名人 |
| 16 | 以前相手に渡した物 |
| 17 | 相手について考慮すること（金属アレルギー，苦手な匂いなど） |
| 18 | 相手が新しく買ったもの |

大学生 A は友人に贈るプレゼントを探すために，商品検索を行う．贈る相手の好きなものを考えて，コスメを贈ることを考える．コスメを検索し，候補を定める．しかし，質問欄に目を向けると，「相手が新しく買ったもの」といった質問が存在したため，その質問について考察を行うことで，最近新しいコスメを購入していたことを思い出し，コスメを候補から外す．次にどの商品が最適か考察していると「相手の持ち物で傷んでいるもの・買い替え時のもの」という質問に目が止まる．再び相手について考察すると，定期入れが長年使用されており傷んでいることを思い出した．このことから，定期入れの検索を行っていくと，相手が好きなキャラクターの定期入れが見つかった．以上の行動から，プレゼントは定期入れに決定した．

上記のように質問を提示することで，相手について考慮することが促進され，プレゼント選択の幅が広くなると考える．以上を踏まえて，表 1 に示した質問項目を用いたシステムを提案する．

提案システムは，Web ページで利用することを想定し，HTML，JavaScript で実装した．本稿では，楽天株式会社が展開する楽天ウェブサービスの楽天商品検索 API (version:2017-07-06) を利用した．図 1 に提案システムのインタフェースを示す． 質問の回答を記入式にし，システム起動時から常時質問を提示する（図 1-A 参照）．ユーザがプレゼントを選択する中で任意の機会に質問の回答を行う．質問をきっかけに検索キーワードを入力する（図 1-B 参照）ことでそのキーワードに該当する商品が表示される（図 1-C 参照）．質問提示部分を質問欄とする．


図 1: 提案システム

## 4 実験

### 4.1 実験の概要

本実験では，プレゼントを選択する過程に着目し，贈り手は提案システムを用いることで相手についての考慮が促されるかを検証することを目的とする．そのため，提案システムから質問欄を除いた商品検索機能のみのシステムを従来型システムとし，提案システムと比較を行う．従来型システムを用いた実験は，情報系学部の大学生 4 名（男性 1 名，女性 3 名）を実験協力者とし，提案システムを用いた実験は，従来型システムを用いた実験に参加していない情報系学部の大学生 4 名（男性 2 名，女性 2 名）を実験協力者とした．

実験協力者には，各々に割り当てられたシステムを用いて，プレゼント選択を行ってもらった．実験協力者に対して述べた実験の条件は，(1) プレゼントを贈る相手は実験協力者自身に決定してもらうが，例外として事前に贈るものやカテゴリを決定している相手の選択は避けてもらうこと，(2) プレゼントを選択する際に，思っていることや考えていることはできるだけ発話すること，(3) 贈るプレゼントが決定した時点で，実験は終了とすること，である．

実験の流れを以下に示す．まず実験協力者に対して実験内容を伝え，実験で用いるシステムの使用方法について説明を行った．次にシステムに慣れるため，指定した検索ワードを実験協力者に入力してもらい，商品検索を行ってもらった．提案システムの説明時，システムの質問欄に関しては，記入できることのみを伝え，質問欄の使用を促すことがないよう配慮した．その後，実験協力者にプレゼントを贈る相手を決定してもらい，実験開始とした．実験は発話思考法で行い，実験協力者の実験中における発言に対し，即座に対応するため，実験協力者の隣には実験者が待機した．実験中，商品検索の過程で出た検索ワードに対し，「なぜそのワードで検索を行ったか」といった質問を行った．これは，プレゼント選択時にどのような理由で検索を行ったかを明確にするためである．実験協力者の同意を得た上で，

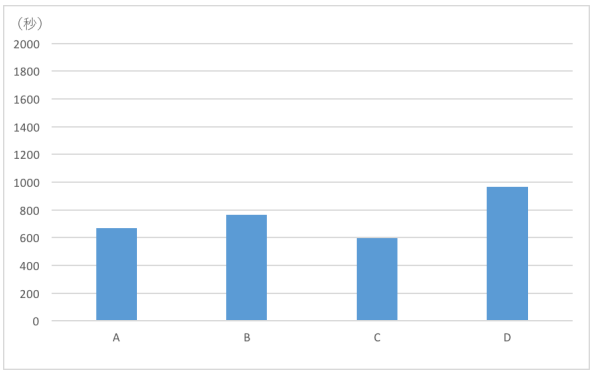



図 2: 従来型システムの検索時間

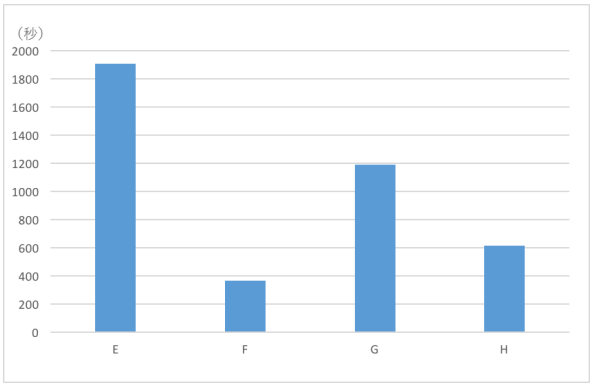



図 3: 提案システムの検索時間

実験中の検索画面の録画と発話の録音を行った．

### 4.2 検索時間とアンケートの結果

従来型システムを用いた実験協力者のプレゼント選択の平均検索時間は 12 分 29 秒であった．一方，提案システムを用いた実験協力者のプレゼント選択の平均検索時間は 17 分 00 秒であった．図 2 に従来型システムを用いた実験協力者の検索時間を示し，図 3 に提案システムを用いた実験協力者の検索時間を示す．

最終的な決め手と悩んだ商品を選択しなかった理由についてのアンケートを行った結果を記す．まず，従来型システムを用いた実験協力者 A，B，C，D のアンケート結果について記す．実験協力者 A は，相手の外見の考慮や値段から最終決定を行った．実験協力者 B は，価格や機能性，目新しさ，見た目から商品を選択した．実験協力者 C は，値段や見た目，相手の好みから商品決定をした．実験協力者 D は，相手が使いやすいかという点で商品を選択した．次に，提案システムを用いた実験協力者 E，F，G，H のアンケート結果について記す．実験協力者 E は，商品が消耗品である点や相手への思い，相手の好みから商品決定を行った．実験協力者 F は，季節や自身の好み，値段を考慮し，商品を選択した．実験協力者 G は，相手への思いや身につけているものから最終決定を行った．実験協力者 H は，実用性や相手の気持ちを考慮した商品決定を行った．

## 5 議論

### 5.1 検索時間の考察

従来型システムと提案システムのそれぞれの実験協力者の検索時間を比較すると，平均検索時間は従来型システムよりも提案システムの方が長い結果となった．しかし，プレゼントを選択する贈り手によって，選択する時間や悩む時間に差があったため有意な差はないと考えられる．実験終了後，実験協力者全員に普段プレゼントを選択する際，時間をかけて選択する方か否かといった質問を行った．普段からプレゼントを選択するのが早いと答えた実験協力者は，実験においても商品を即決する様子が観察された．また，同じ提案システムを用いた実験協力者 E と F を見てみると， 25 分の時間差があることがわかる．実験協力者 E は提案システムの質問欄に全て目を通し，回答する様子が発話から観察されたのに対し，実験協力者 F は質問欄を使用することなく，商品決定に至ったことから時間差が生まれたと考えられる．

### 5.2 検索過程の分析

まず，データを読み込むために，実験から得られたプレゼント選択過程の音声データをテキストとして文字に起こした．次に，その記録された発話を文単位で分割した．最後にそれらを元に内容を簡潔に表すラベル名をつけ，類似しているラベルごとにカテゴリとしてまとめた．例を表 2 に示す．従来型システムと提案システムにおいてそれぞれのカテゴリを相手についての考慮を含むカテゴリ（e.g., 相手の趣味）と相手についての考慮を含まないカテゴリ（e.g., 商品の発見）に分類した．これは，検索過程において相手への考慮が促進されたかを比較するためである．分類したカテゴリのうち，相手についての考慮を含むカテゴリを表 3 に示す．

従来型システムでは，相手についての考慮を含むカテゴリが 12 個生成された．一方，提案システムでは，カテゴリが 17 個生成された．両システムともに得られたカテゴリとしては相手の状況，イベント，趣味，身につけているもの，エピソード，好きな色，好きな芸能人，好み，以前買ったものの想起，考慮の 9 個である．これらから，両システムともに相手についての考慮がみられたことがわかる．しかし，提案システムを

表 2: カテゴリ生成例

| カテゴリ | 分割した文 | ラベル |
|---|---|---|
| 相手の状況を考慮 | 内定貰った企業で基本情報のなんか資格を取らされるとのことだったので，じゃあそれに関する参考書にしましょう（基本情報技術試験　参考書で検索） | 相手の内定先を考慮 |
| | これから稼ぐってね，まだそんなお金持ちじゃないから価格は低めの方がいいかな | 相手の状況を考慮 |
| | （商品説明を見て）横向きに寝たときの寝姿勢が安定します，あー A 先生がどっち向いて寝るとか分かれへん | 商品説明を確認し，相手について考察 |

表 3: 相手についての考慮を含むカテゴリ

| システム | カテゴリ |
|---|---|
| 従来型システム | 相手の状況を考慮<br>相手のイベントの考慮<br>相手の趣味を想起<br>相手が身につけているものの想起<br>相手のエピソードを想起<br>相手の好きな色の想起<br>相手の好きな芸能人を考慮<br>相手の好みの考慮<br>相手が以前買ったものの想起<br>相手の外見を想起<br>相手の居住地の想起<br>相手の癖から想起 |
| 提案システム | 相手の状況を考慮<br>相手のイベントの考慮<br>相手の趣味を想起<br>相手が身につけているものの想起<br>相手についてのエピソードを想起<br>相手の好きな色の想起<br>相手の好きな芸能人を考慮<br>相手の好みの考慮<br>相手が以前買ったものの想起<br>好きなブランドの想起<br>相手に以前渡したものの想起<br>相手が傷んでいるものや買い替え時のものの想起<br>相手の好きなキャラクターの想起<br>相手の好きな食べ物の想起<br>相手が大切にしているものの想起<br>相手が好きなテレビの想起<br>相手の考慮することの想起 |

用いた場合のみ得られた 8 個のカテゴリは，システムが提示を行った質問と同一の項目である．実験協力者の発話から提案システムを用いた 4 名中 3 名がシステム上で提示した全ての質問を確認し，その質問をきっかけに商品検索を行ったことが確認された．そのため，システム上で質問を提示したことによりそれらのカテゴリが得られたと考えられる．また，カテゴリの数からシステム上で質問を提示することで，提案システムは従来型システムに比べ，多様な観点から相手について考慮させていたことが示唆される．

表 4: 従来型システムの相手について考慮した回数

| 実験協力者 | 回数 |
|---|---|
| A | 6 |
| B | 8 |
| C | 6 |
| D | 5 |
| 平均 | 6.3 |

表 5: 提案システムの相手について考慮した回数

| 実験協力者 | 回数 |
|---|---|
| E | 36 |
| F | 1 |
| G | 39 |
| H | 18 |
| 平均 | 23.5 |

## 5.3　相手について考慮した回数の考察

両システムの実験協力者の発話から相手について考慮した発話回数を測るため，カテゴリ生成の際に分割した文ごとに相手についての考慮が行われた発話回数を計測した．その結果を表 4，5 に示す．従来型システムを用いた実験協力者は平均 6.3 回，提案システムを用いた実験協力者は平均 23.5 回となった．また提案システムを用いた実験協力者のうち，2 名は相手について考慮した発話が 30 回以上見られた．このことから，提案システムを用いることで，相手について考慮する発話が増えたため，相手の考慮が促されたことが示唆される．

## 5.4　アンケート結果の考察

プレゼントの最終的な決め手についてアンケートを行った結果を述べる．相手について考慮した回答（e.g., 相手の好みや使いやすさ）が従来型システムでは 4 名中 3 名から，提案システムでは 4 名全ての実験協力者から得られた．一方で，相手について考慮されていない回答（e.g., 価格や商品の見た目）が従来型システムでは 4 名中 3 名から，提案システムでは 4 名中 2 名から得られた．このことから両システムともに相手に

ついて考慮したプレゼント選択が行われていたことが確認された．さらに，悩んだ商品を選択しなかった理由については，最終的な決め手と同様に，両システムともに相手について考慮した結果が得られた．これはプレゼント選択において，相手について考慮することは必要なことであるため，両システムにおいて最終的な決め手や悩んだ商品を選択しなかった理由に相手を考慮した結果が反映されたと考えられる．

### 5.5 全体の考察

本稿では，プレゼント選択の中で相手についての質問を提示することで，相手について考慮する行為を促進させることを目的とした実験を行った．アンケート結果から，両システムともに相手への考慮が観察された．検索過程から相手への考慮がどれほど促されたかを観察した結果，提案システムを用いることで相手についての考慮を含むカテゴリが多く得られた．また，相手について考慮した発話回数を計測した結果，提案システムでは従来型システムに比べ，より多くの発話が観察された．これらから，システム上で相手についての質問を提示することにより，多様な観点から相手について考慮することができたと考えられる．しかし，提案システムでは従来型システムを用いた場合のみ考慮されていた相手の外見，居住地，癖といったカテゴリが得られなかった．このことから質問項目を検討する必要があると考えられる．

## 6 おわりに

本稿では，プレゼント選択の過程において，贈り手が相手についての考慮を促すことで，納得した商品選択を行うことを目的とした．相手についての質問を提示した提案システムを用いて，相手への考慮が促進されるかを目的とした実験を行った．その後実験協力者の発話から分析を行ったところ，提案システムを用いた場合の方がより多様な観点から相手への考慮がみられ，また相手について考慮する発話が増えた．今後は，贈り手に対し，より多様な観点から相手についての考慮を促すため，質問項目の追加と改良を検討していく．

## 謝辞

本研究の実施にあたり JSPS 科研費 15H02780 の助成を受けた．記して謝意を表す．

## 参考文献

[1] Breese, J. S., Heckerman, D. and Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 43–52 (1998).

[2] Linden, G., Smith, B. and York, J.: Amazon.com Recommendations: Item-to-item Collaborative Filtering, *IEEE Internet computing*, Vol. 7, No. 1, pp. 76–80 (2003).

[3] 空中海人, 上間大生, 松下光範: オンラインショッピングにおける内省行為に着目した贈り物選定の支援, Web インテリジェンスとインタラクション研究会予稿集, No. 4, pp. 81–86 (2014).

[4] 盆子原健太, 大塚直也, 松下光範: EC サイトにおける商品探索プロセスに着目したプレゼント探索支援システムの提案, 第 6 回インタラクティブ情報アクセスと可視化マイニング研究会, SIG-AM-06-06, pp. 34–39 (2014).

[5] 田口拓明, 田村哲嗣, 速水悟: 商品レビューを用いたプレゼント支援の検討, 2014 年度人工知能学会全国大会（第 28 回）論文集, 3M3-3 (2014).

[6] 村上知子, 吉岡信和, 折原良平, 古川康一: CAM 法を用いた個人嗜好モデルに基づく商品推薦システム, 人工知能学会論文誌, Vol. 20, No. 5, pp. 346–355 (2005).

[7] 水野淳太, 村田祐一, 勝屋久: ユーザの嗜好を反映したクエリ拡張を用いた情報検索・推薦システムの開発, 楽天研究開発シンポジウム (2009).

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第18回)
SIG-AM-18-09

# キャラクタの出現頻度に着目したコミックのエピソード分割手法の検討

# Episode division method in comics by using frequency of personages' appearance

中本 竣也 [1]　朴 炳宣 [2]　松下 光範 [1*]
Shunya Nakamoto[1]　Byeongseon Park[2]　Mitsunori Matsushita[1]

[1] 関西大学 総合情報学部
[1] Faculty of Informatics, Kansai University

[2] 関西大学大学院総合情報学研究科
[2] Graduate School of Informatics, Kansai University

**Abstract:** This paper propose a method for dividing story in a comic by episode. In recent years, it has been possible to search comics by using bibliographic information such as publishers and authors. It, however, is still difficult to search the comics by focusing on the contents of them. To solve the problem, we aim to develop a system to search an intended episode of a comic. To achieve this goal, we focus on the frequency of personages' appearance. This idea is based on hypothesis that the trend of personages appearance relates to the episode's progress in comics. Our experiment so far revealed that weighting the importance of personages (e.g., main character and rival character) can contribute to divide episodes correctly.

## 1　はじめに

電子書籍の普及，発展に伴って Web 上やタブレット端末，およびスマートフォンなどのディジタルデバイスで読むことができるディジタルコミックが登場しており，他の電子書籍と同様に利用者数が増加している．電子書籍市場の 8 割以上をディジタルコミックが占めており，その発展に大きく貢献している [8]．コミックの電子化によって，従来のコミックでは考えられなかった新しい利用や表現が期待されている．その一つとして，エピソードごとの検索が挙げられる．

電子書籍販売サイトによってコミックの検索が可能であり，これらに対する読者の利用意向は高い [8]．電子書籍販売サイトでコミックの検索を行う場合，作品のタイトルや著者など書誌情報を用いた検索が可能である．一方で，従来のサービスでは，コミックの内容に関する検索はジャンルによる簡略的なものに留まっており，現状では「ルフィがゾロと出会うエピソードを閲覧したい」といったような読者の詳細な要求に応えることができていない．この問題を解決するために，本研究ではコミックから特定のエピソードを検索するための技術の実現を目的とし，その端緒として，コミックに書かれた内容情報に基づきエピソードを自動的に分割する手法を検討する．エピソードはコミックのページ上のコマ割りとそのページの連続から表現されるものであり，複数のエピソードによりコミックが構成されるという特徴がある [10]．コミックが出版される際，エピソードは明示されない事が多く，エピソードを明確に分割することが難しい．コミックは出版形式（e.g., 週刊・月刊誌，単行本）による単位（e.g.,「巻」「話」）をもとにエピソードが構成され，コミックを分割する際，現状では，巻数や話数が考えられる．しかし，必ずしも一つの単位ごとに一つのエピソードが表現されているとは限らず，複数の「巻」や「話」に渡って一つのエピソードを構成する場合がある．そこで本研究では，コミックに含まれた内容情報（e.g., キャラクタ，セリフ）からエピソードを特定し，それらを分割することを試みる．

*連絡先：関西大学総合情報学部総合情報学科
〒 569-1095 大阪府高槻市霊山寺町 2-1-1
E-mail: mat@res.kutc.kansai-u.ac.jp

## 2　関連研究

### 2.1　ストーリ推定に関する研究

これまで様々なコンテンツにおいて，コンテンツを効率的に利用・選択する必要性が高まっており，ダイジェストや要約表現の自動生成に関する研究が国内外で行われている．テキスト要約に関する研究は盛んに行われているが，ストーリ抽出においては，自動化できていない現状である．相良らの研究では，テキストのストーリはメインストーリとサブストーリの組み合

わせから構成されるという考えに基づき，従来の重要文抽出を利用したテキストからのストーリ抽出の手法が提案され，有効性について確認されている [7]．岩永らの研究では，野球の試合に関する報道を取り上げ，試合中に発生したイベントについて報じたテキストを入力とし，それらを基にその試合のダイジェストを生成する手法について提案している [6]．各イベントは試合中に生じた全てのイベントの通し番号，イニング，裏表，攻撃側のチーム名，打者，打席の内容，盗塁，選手交代などからなる．吉高らは，映画やドラマなどの映像コンテンツを対象とし，製作者がある場面を演出するための技法に着目する手法を提案した [11]．映像や音声に現れる特徴に着目することで，効果的なダイジェスト生成を目指している．印象的な場面であるかどうかの判断は視聴者によってばらつきがあると考えられるが，このような手法から制作する側が意図的に作った印象的な場面を検出することが可能であると示唆された．

### 2.2 内容情報の自動識別に関する研究

コミック内のキャラクタの自動識別に関する研究や，コミックの自動コマ分割に関する研究は盛んに行われている [1][4]．石井らの研究では，コミックのコマ分割処理において，帯を用いた分割線候補の検出，および分割線適合検査を用いる手法が提案されている [4]．さらに，分割の際に用いる閾値の最適化を図り，分割線検出精度の向上を確認している．野中らはスマートフォンのような小画面の端末における電子コミックの閲覧を容易にするため，電子コミックのコマ検出を自動的に行うことを可能にする画像解析技術 GT-Scan を開発した [9]．GT-Scan は複数の画像処理モジュールの組み合わせによって構成され，入力されたコミック画像を解析してコマ情報を出力する．GT-Scan によってコマ情報付与の作業時間を約 70%削減することができ，また少年漫画では 95%，少女漫画では 78%の精度でコマの自動検出が行われていた．それに付随し，コマ情報が付与された電子コミック用データをスマートフォンで閲覧する際に，コマ単位に閲覧できる GT-Comic Viewer を開発，商品化し，読みやすい電子コミックを低コストで提供可能な環境を整えた．

### 2.3 コミックへのアクセス支援に関する研究

現在，コマの同定 [3] や，構成要素に基づく自動シーン分割処理 [5]，書誌またはコンテンツ情報の構造化 [2] などに関する研究が行われている．石井らの研究では，各コマに含まれる構成要素の分布に対する解析から，現時点で自動取得可能とされるメタデータではナレーション

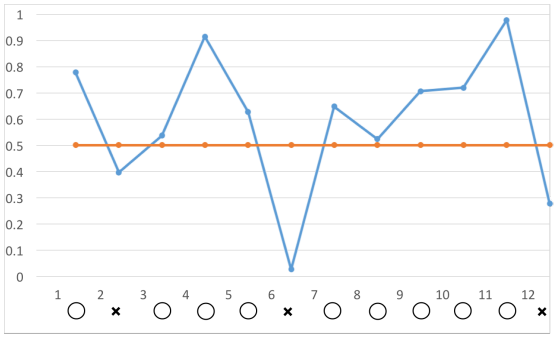



図 1: 提案手法による区切り位置

が，現時点で自動取得が困難なメタデータでは背景の距離がシーン切り替わりにおいて最も影響力が大きい要素であると明らかにされた [5]．Morozumi らは，ネットワーク環境上でのコミックへのアクセスや，再利用のためのメタデータフレームワークを提案している [2]．

## 3 提案手法

コミックを構成する要素の中でキャラクタは最も頻繁に出現する要素であり，最も重要な要素である [10]．エピソードの区切りでは，そのエピソードに特有のキャラクタの登場や退場が頻繁に行われている．例えば,「ある敵を倒すまでの物語」というエピソードでは，敵が登場した部分や敵が退場した部分がエピソードの区切りとなる．そこで本稿では,「キャラクタの登場や退場はエピソードの区切りを表す重要な要素である」と仮定し，キャラクタの出現頻度に着目する．同じ登場人物が共通して出現している話同士は，一つのエピソードに分類されると考えられ，出現している登場人物が異なる話同士は異なるエピソードに分類されると考えられるため，本稿の提案手法では，一話ごとのキャラクタの出現頻度を用いて隣接話間の類似度を測り，エピソード分割の指標とする．

本稿では類似度の算出に cos 類似度を用いた． cos 類似度は，二つの $n$ 次元のベクトル間の距離を測る際に用いられ，値が 1 に近いほど二つのデータが似通っていることを示す． cos 類似度を求める数式を以下に示す．

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}\,|\,|\vec{d}\,|} \tag{1}$$

出現頻度に基づくエピソードの区切り位置の判別例を図 1 に示す．図 1 の「×」は該当箇所が 区切りであることを示している．なお本稿では指標によって得られた隣接話間の類似度が 0.5 未満の場合をエピソードを区切る箇所とした．例えば，図 1 の場合，2 話と 3 話，

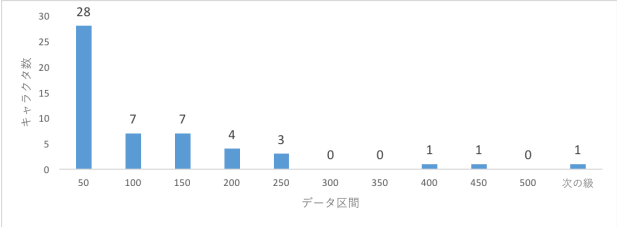



図 2: キャラクタの出現頻度分布 (進撃の巨人)

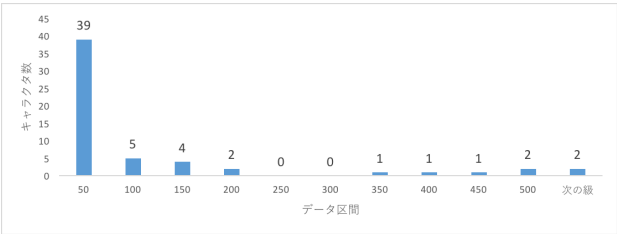



図 3: キャラクタの出現頻度分布 (DEATH NOTE)

6 話と 7 話，12 話と 13 話の間の 3 箇所がシステムがそれぞれのエピソードの区切り位置であると提示していることを表す．

# 4 予備実験

## 4.1 予備実験の概要

ベースラインとして各話のキャラクタの出現回数をベクトルで表現し，隣接話間の cos 類似度を求めた（以下，指標 (A) と呼ぶ）．使用したコミックは『銀魂』(空知英秋著，集英社),『進撃の巨人』(諫山創著，講談社)，『DEATH NOTE』(大場つぐみ，小畑健著，集英社)，『焼きたて!!ジャぱん』(橋口たかし著，小学館）を対象とした．指標 (A) によって提示されたエピソード分割と正解データの比較実験を行い，エピソード分割の一致率を算出した．

表 1: 指標ごとの一致率（%）

| 作品名 | A | B | C |
|---|---|---|---|
| 銀魂 | 28.2 | 92.3 | 82.1 |
| 進撃の巨人 | 47.6 | 66.7 | 66.7 |
| 焼きたて!!ジャぱん | 57.5 | 67.5 | 60.0 |
| DEATH NOTE | 71.4 | 59.5 | 69.0 |
| 平均 | 51.2 | 72.6 | 69.5 |

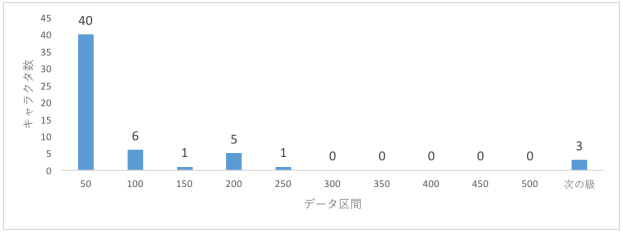



図 4: キャラクタの出現頻度分布 (銀魂)

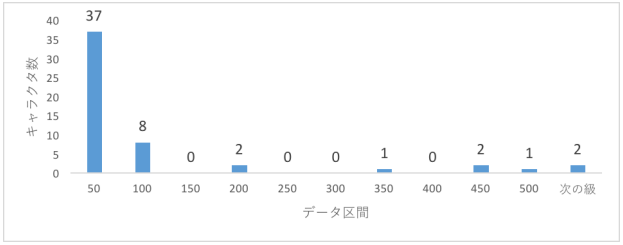



図 5: キャラクタの出現頻度分布 (焼きたて!!ジャぱん)

## 4.2 予備実験の結果

表 1 の A 列に示すように，指標 (A) からは十分な精度が得られなかった．原因として，いずれの作品についても高い類似度が示される箇所が多く，0.5 を下回る箇所が少なかたったため，区切りとして検出される箇所が少ないことが挙げられる.『DEATH NOTE』と『焼きたて!!ジャぱん』において，特に多くの箇所で類似度が高いことが確認された．これはある特定の箇所では出現に変化があるキャラクタ数より，変化がないキャラクタ数が多いことが原因であると考えた．そこで，4 作品について，キャラクタの出現頻度の分布を調査した．その結果を図 2〜図 5 に示す.『進撃の巨人』と『銀魂』の 2 作品では 300 回以上出現していたキャラクタは 3 キャラクタのみだったが，特に高い類似度が示される話が多かった『焼きたて!!ジャぱん』と『DEATH NOTE』の 2 作品は 300 回以上出現していたキャラクタは，それぞれ 6 キャラクタ，7 キャラクタ存在していた．このことから,『DEATH NOTE』と『焼きたて!!ジャぱん』は同じキャラクタによって構成される話が多いため，特に高い類似度が提示される箇所が多かったと考えられる．

常に出現しているキャラクタに比べ，稀に出現するキャラクタが各エピソードの特徴となると考えられるため，キャラクタの出現の傾向に対する重み付けを行うことで，実態を反映した類似度を得ることが期待される．そこで実験では指標 (A) に加え，2 種類の重み付け手法について検討することでエピソード分割の精度向上を図った．

# 5 実験

本稿の提案手法によってエピソードごとの分割を行うにあたり，キャラクタごとに重み付けを行うことでエピソード分割の精度向上が期待される．提案手法の有効性の確認，適切な重み付け手法の検討のために，それぞれの精度を比較する実験を行った．

## 5.1 実験の概要

指標 (A) に加え，2 種類の重み付け手法について検討した．本実験では，予備実験と同様に『銀魂』,『進撃の巨人』,『DEATH NOTE』,『焼きたて!!ジャぱん』の 4 作品の 1 巻〜5 巻を対象とした．

一つ目の重み付け手法は TF-IDF の考え方を用いたものである．ある特定のエピソードにのみ出現するキャラクタは，そのエピソードを特徴づけるキャラクタであると考えられるため，話ごとのキャラクタの出現回数を TF とし，各話に対する出現回数の逆数を IDF とした．それにより，特定のエピソードに含まれる話に多く出現し，他の話では出現が少ないキャラクタにより高い重みを与える．この数式を以下に示し，この重み付けを指標 (A) に反映したものを指標 (B) とする．

$$w_{ij} = \mathrm{TF}_{ij} \cdot \mathrm{IDF}_i \tag{2}$$

$$\mathrm{TF}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{3}$$

$$\mathrm{IDF}_i = \log \frac{C}{df_i} \tag{4}$$

ここで，$n_{ij}$ はある $j$ 話内におけるキャラクタ $i$ の出現回数であり，$\sum_k n_{kj}$ は $j$ 話におけるすべてのキャラクタの出現回数の和である．$C$ がコミックの総話数，$df_i$ は，あるキャラクタ $i$ が出現する話の数を示している．

二つ目の指標は，キャラクタの出現間隔を用いた重み付け手法である．全ての話に渡って連続的に出現するキャラクタよりも，非連続的に出現していたキャラクタが再度出現する話は，その直前のエピソードと異なるエピソードに属すると考えられる．そこで，対象キャラクタが直前に出現してから次に出現するまでの間隔を重みとし，指標 (A) に反映したものを指標 (C) とする．数式を以下に示す．

$$w_{ij} = \frac{n_{ij}}{p_j} \log 2N_i \tag{5}$$

$n_{ij}$ はある $j$ 話内におけるキャラクタ $i$ の出現回数であり，$p_j$ は $j$ 話の総ページ数，$N_i$ はキャラクタ $i$ が直前に出現してから次に出現するまでの間隔を，それぞれ表している．

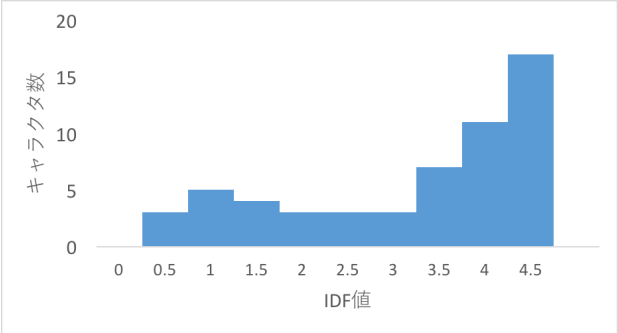



図 6: IDF 値の分布（銀魂）

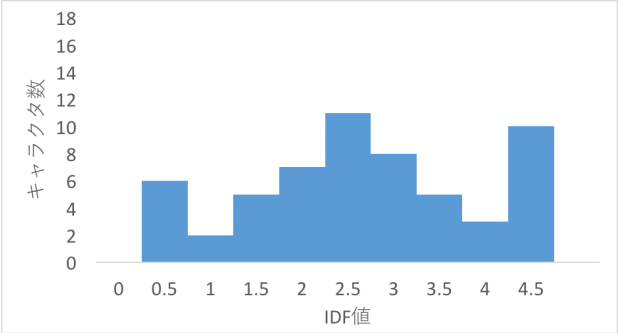



図 7: IDF 値の分布（DEATH NOTE）

## 5.2 実験結果

指標 (B)，(C) を用いて得られたデータと人手による正解データの比較結果を表 1 の B 列， C 列に各々示す.『銀魂』,『進撃の巨人』,『焼きたて!!ジャぱん』の 3 作品に対して (B) と (C) どちらの指標を用いた場合も (A) よりも高い正答率が示された．特に『銀魂』において (B) の指標を用いた際に，92.3%の精度でエピソードを正しく分割できた．このことから，キャラクタごとに重み付けを行った上で類似度を求める手法によるエピソード分割の有効性が示された．一方,『DEATH NOTE』では (B) と (C) のどちらの指標を用いた場合も，(A) より低い正答率が示された．

# 6 考察

指標 (B)，(C) によって，一致率の向上が確認されたため，キャラクタごとに出現の傾向に基づき重み付けを行った上で，隣接話間のキャラクタの出現頻度について類似度を測り，エピソード分割の指標とする手法の有効性が示された.『銀魂』は特定の話にのみ出現しているキャラクタが存在し，このようなキャラクタの IDF 値が常に出現するキャラクタ（e.g., 主人公）よりも高かった．これにより，キャラクタの登場と退場によって各エピソードの分割が可能であることが示唆さ

れた．一方,『DEATH NOTE』は出現するキャラクタ数が少なく，エピソードの変化をキャラクタの種類ではなく，少ないキャラクタによるセリフや行動で表現している場合が多く見受けられた.『DEATH NOTE』と『銀魂』の IDF 値の分布を調査したところ,『DEATH NOTE』は『銀魂』に比べ，IDF 値が低いキャラクタの数が少なかった（図 6，図 7 参照）.『DEATH NOTE』は『銀魂』に比べ，キャラクタごとの IDF 値の差が小さく，各エピソードの特徴となるようなキャラクタの数が少なかったことが原因と考えられる．

## 7　おわりに

書籍販売サイトなどではコミックの書誌情報を用いた検索が可能になっている．一方で，コミックの内容情報による検索はジャンル検索のような簡略的なものに留まっている．そのため，本研究ではコミックからエピソードの検索を行う技術の実現を目的とし，その端緒として本稿では，コミックに混在する要素の中からキャラクタの出現頻度に着目した．コミックをエピソードごとに自動的かつ定量的に分割する手法を試みた．人手による分割と提案手法による分割の比較実験を行った結果，キャラクタごとに重み付けを行った上で類似度を求める手法におけるエピソード分割の有効性が示された．

『DEATH NOTE』のような作品は，出現するキャラクタではなく，キャラクタの状態や様子，言動により各エピソードを表現している場合が多かった．そのため，このようなコミックについては本稿の手法では十分な精度が得られなかった．今後は，このようなコミックに対して，キャラクタの出現頻度以外のコミックを構成する要素に着目したエピソード分割手法を検討する必要がある．また，コミックを分類することができればコミックごとに適切なエピソード分割手法を選択できると考えられる．

また，現在では，コミックやコミックを原作とするアニメーション作品のダイジェストを自動生成する研究は行われているが，本研究で用いた手法でエピソード分割が可能になった場合，ダイジェストの生成に応用できると期待される．

## 謝辞

本研究の実施にあたり JSPS 科研費 15K12103 の助成を受けた．記して謝意を表す．

## 参考文献

[1] Ishii, D., Yamazaki, T. and Watanabe, H.: Multi Size Eye Detection on Digitized Comic Image, *IIEEJ 3rd Image Electronics and Visual Computing Workshop*, pp. 1–4 (2012).

[2] Morozumi, A., Nomura, S., Nagamori, M. and Sugimoto, S.: Metadata Framework for Mnga: A Multi-paradigm Metadata Description Framework forDigital Comics, *International Conferenceon Dublin Core and Metadata Applications*, pp. 61–70 (2009).

[3] Tanaka, T., Shoji, K., Toyama, F. and Miyamichi, J.: Layout analysis of tree structured scene frames in comic images, *20th International Joint Conference on Artificial Intelligence*, pp. 2885–2890 (2007).

[4] 石井大祐, 河村圭, 渡辺裕: コミックのコマ分割処理に関する一検討, 電子情報通信学会論文誌, Vol. J90-D, No. 7, pp. 1667–1670 (2007).

[5] 石井大祐, 柳澤秀彰, 三原鉄也, 永森光晴, 渡辺裕: マンガの構成要素に基づく自動シーン分割処理に関する一検討, 電子情報通信学会技術研究報告, Vol. 114, No. 349, pp. 73–76 (2014).

[6] 岩永朋樹, 西川仁, 徳永健伸: テキスト速報を用いた野球ダイジェストの自動生成, 言語処理学会発表論文集, No. 22, pp. 238–241 (2016).

[7] 相良直樹, 砂山渡, 谷内田正彦: 重要文抽出を利用したテキストからのストーリー抽出, 情報処理学会研究報告自然言語処理. 2004-NL-164, No. 108, pp. 159–164 (2004).

[8] 電子書籍ビジネス調査報告書 2016: インプレス総合研究所 (2016).

[9] 野中俊一郎, 沢野哲也, 羽田典久: コミックスキャン画像からの自動コマ検出を可能とする画像処理技術「GT-Scan」の開発, *Fuji Film RESEARCH & DEVELOPMENT*, No. 57, pp. 46–49 (2012).

[10] 三原鉄也, 永森光晴, 杉本重雄: ディジタルコミックにおけるストーリ構造とビジュアル構造を表すメタデータモデル, 情報処理学会研究報告. FI, No. 9, pp. 1–8 (2011).

[11] 吉高淳夫, 田中壮詩, 平嶋宗: 映画等を対象としたダイジェスト映像生成のための映像特徴に関する検討, 情報処理学会研究報告, Vol. 2007, No. 68, pp. 79–86 (2007).