Vol.2021-MBL-100 No.6
Vol.2021-UBI-71 No.6
Vol.2021-CDS-32 No.6
Vol.2021-ASD-21 No.6
2021/9/2

類似生活者を発見するトランザクションデータベース距離

杉村 博[1,a)]　酒井 貴洋[1,b)]　一色 正男[1,c)]　松本 一教[1,d)]

概要：本論文では 2 つのトランザクションデータベース間の距離計算手法を提案し，スマートホームのデータから類似の行動をする生活者を発見するシステムを構築する．生活者の行動による家電の動作状況をアイテムとして，複数のアイテムが同時共起するトランザクションの集合であるトランザクションデータベースを構築する．各生活者のデータベースが与えられるとき，それらの差を距離として計算することによって，類似の生活者を発見可能となる．単純なアイテムの出現頻度や相関ルールマイニングといった各データベースの特徴をもとにして，2 つのトランザクションデータベース間の距離を計算する方法とその実験結果について説明する．

キーワード：距離計算手法, トランザクションデータベース, スマートホーム

The Distance of Transaction Databases to Discover Users with Similar Lifestyle

Hiroshi Sugimura[1,a)]　Takahiro Sakai[1,b)]　Masao Isshiki[1,c)]　Kazunori Matsumoto[1,d)]

Abstract: This paper proposes a distance calculation method between two transaction databases and constructs a system to discover users with similar lifestyles. Using the operation status of home appliances by the consumer's behavior as an item, the smart-home system constructs a transaction database that is a collection of transactions in which multiple items co-occur at the same time. When giving multiple databases, it is possible to judge similar consumers by calculating their differences as distances. We describe the method for calculating the distance between two transactions based on the features of each database, such as the frequency of occurrence of items or correlation rule mining.

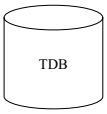
Keywords: Distance method, Transaction database, Smart home

1. 序論

本研究はスマートホームのデータベース（以下，DB）解析として，類似生活者を発見することを目的とし，複数のトランザクションデータベース（以下 TDB）間の距離を計算するアルゴリズムを提案する．スマートホームの TDB に記録されるデータの種類は多岐に渡り，家族構成のような静的データ，時間変化を記録する時系列データ，センサからの数値データ，家電の状態を記録するカテゴリデータ等が記録される．今後のスマートホームの機能には，これらのデータを収集して分析し，生活を支援する仕組みが期待される．データ分析手法には，統計学的なものから人工知能的なデータマイニング技術等があり，目的に応じて利用する手法が異なるが，本研究の目的は類似生活者を発見して，グルーピングする手法を目指している．類似の生活者が発見できれば，数か月後や数年後の未来状態予測や，健康維持行動のレコメンド等に活用可能である．更に，グルーピング結果及び計算モデルからスマートホームサービスを構築するための知識抽出を行うことを念頭にしている

[1] 神奈川工科大学
Kanagawa Institite of Technology, Atsugi, Kanagawa 243–0292, Japan
[a)] sugimura@he.kanagawa-it.ac.jp
[b)] tsakai20@ele.kanagawa-it.ac.jp
[c)] masao@he.kanagawa-it.ac.jp
[d)] matumoto@ic.kanagawa-it.ac.jp

Vol.2021-MBL-100 No.6
Vol.2021-UBI-71 No.6
Vol.2021-CDS-32 No.6
Vol.2021-ASD-21 No.6
2021/9/2



トランザクションID	アイテム集合
1	{ ビール, 焼き鳥, タバコ }
2	{ おにぎり, お茶, 焼き鳥, 新聞 }
3	{ 新聞, おにぎり, お茶, タバコ }

図 1 トランザクションデータベース
Fig. 1 Transaction Database.

ため，ディープラーニング [1] のような，出力モデルを人間が解析することが困難なアルゴリズムは適していない．

提案する距離計算アルゴリズムは，スマートホーム 1 軒を 1 つの TDB と考え，2 つの TDB 間の違いを距離として算出する．例えば 2 組以上の TDB 間の違いを距離として算出すれば，一方の距離を基準とし，他方の距離が近いか遠いかを相対的に評価できる．更にその距離がどのような特徴にもとづいて算出されたかも保存しているので，TDB 間の距離の理由を発見することも可能になる．TDB 間の距離が計算できれば，グルーピングは従来の階層型クラスタリング [2] や k-means[3] 等のクラスタリングアルゴリズムを単純に適用可能となる．

TDB の性質と，従来の TDB からの知識発見手法について紹介する．TDB は一連の処理をトランザクションという単位で蓄積して管理する DB である．元来のトランザクション処理は複数の DB 操作の一貫性を確保する仕組みであるが，ここから派生して TDB は複数のアイテムデータの同時生起を重視するデータを扱う DB であることを示すミームになっており，本研究でもこのように表現する．例えばスーパーマーケットでの商品販売記録は 1 回の商品取引において複数の商品を同時に販売しており，各商品をアイテム，1 回の取引を 1 トランザクションとして記録する TDB である．

TDB からの知識発見としては頻出アイテムやその集合，相関ルールの抽出が良く知られている [4], [5]．また，相関ルールを利用した TDB の分析方法もある [6]．頻出アイテムやその集合の発見は単独の TDB の特徴として，頻出するアイテムやその組み合わせを集合として発見してリストアップするものである．相関ルールの発見は，頻出アイテム集合に対して，更に集合の中でアイテム間の関係性を導くというものである．TDB の特徴としてのみならず，アイテムとアイテムの関係性を導くことができ，データ分析の詳細度が上がっていると見ることができる．いずれにせよこれらのアルゴリズムは，単独の TDB からの特徴抽出を目的としており，複数の DB との違いや類似性の発見に着目した技術はまだない．

2. 従来のデータ間距離

従来の複数データの類似度や距離の計算手法として多様な手法がある [7]．ここでは数値データ間の距離計算手法と，数値以外のデータ間の距離計算手法について紹介し，単純に TDB の距離計算に適用できない理由について説明する．2 つの N 次元ベクトル間の距離としてユークリッド距離（式 1），マンハッタン距離（式 2），Dynamic Time Warping（式 3）等がある．しかし，スマートホームの TDB で取り扱うデータは数値のみならず，家電操作記録のようなカテゴリデータも多数存在するため，数値ベクトルデータの距離計算手法は適用出来ない．

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{1}$$

$$D(x,y) = \sum_{i=1}^{n}|x-a| \tag{2}$$

$$D(x,y) = \min\begin{cases} D(x_i, y_{j-1}) + q & \text{時間ずれコスト q} \\ D(x_{i-1}, y_j) + r & \text{時間ずれコスト r} \\ D(x_{i-1}, y_{j-1}) + s & \text{値不一致コスト s} \end{cases} \tag{3}$$

数値データ以外の距離計算手法として，2 つの文字列の差を示す編集距離（ハミング距離 [8] やレーベンシュタイン距離 [9]）がある．計算方法は Dynamic Time Warping とほぼ同様に，定められたズレや不一致コストをもとにした動的計画法によって計算できる．これら計算手法は文字の出現順序が重要であるため，TDB のようにトランザクションの出現頻度や，一つのトランザクションの出現頻度が重要でない DB に適用することは意味がない．例えば，朝 TV，炊飯を実施したトランザクションと炊飯，朝 TV を実施したトランザクションの編集距離を求めることには意味がなく，出現確率のような特徴を利用するべきである．

アイテム集合の類似度を計算する手法として，Jaccard 距離（式 4）やコサイン類似度（式 5）がある．Jaccard 距離は文書の距離でも使われており，単語の出現共起率で 2 つの距離を計算する．この手法はトランザクションとアイテムの関係においては都合が良いので利用できるが，複数トランザクションと複数トランザクションの対応付けには利用出来ない．コサイン類似度は 2 つのベクトルの類似度を測る手法で，シンプルなレコメンドシステムの協調フィルタリングでよく利用される．この手法はアイテムの有無を 1, 0 を利用してベクトルにエンコードすることでアイテム集合同士の類似度を計算する．この手法も Jaccard 距離のようにトランザクションをベクトルとすれば利用できるが，複数トランザクションに対応できるベクトルエンコード方式が無いので単純には適用できない．

$$D(x,y) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{4}$$

$$D(x,y) = \frac{x \cdot y}{|x| \times |y|} \tag{5}$$

Vol.2021-MBL-100 No.6
Vol.2021-UBI-71 No.6
Vol.2021-CDS-32 No.6
Vol.2021-ASD-21 No.6
2021/9/2

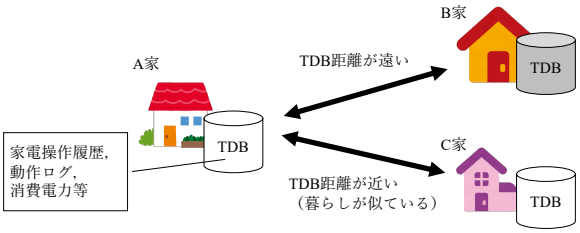


図 2 TDB 間距離からの知識発見

Fig. 2 Knowledge discovery from Distance of Transaction Database.

3. TDB 間距離の提案

本研究では複数の TDB があると仮定して，その TDB 同士の類似性を距離として計算する事を提案する．例えばスマートホームにおいて家電の操作履歴や動作ログ，各家電の消費電力といった生活行動がトランザクションとして蓄積できている場合を考えると，家 1 軒に対して 1 つの TDB としてみなすことができ，Fig. 2 のように TDB 同士で類似性を計算することができれば類似の生活者を発見できる．他にも，家 1 軒の TDB を，1 年毎にサブ TDB として分割すれば，生活行動の年毎の変化度合いを発見することもできると考えた．そこで，アイテムが同時並行的に時系列で記録された TDB が複数あるとき，この TDB 同士の違いを定量的に評価するための TDB 間距離を提案する．

2 つの TDB（TDB_α，TDB_β）間の距離を計算する方法は次の手順で行う．ここでの特徴には，頻出アイテム集合や相関ルール等の，TDB を対象とした特徴獲得アルゴリズムが汎用的に利用できる．本研究では実験的に頻出アイテム集合と相関ルールを用いた手法について具体化して実験を行っている．

(1) TDB_α から抽出された特徴集合全体を F_α とし，TDB_β から抽出された特徴集合全体を F_β とする．

(2) 共通の特徴集合全体を $F = \{F_\alpha \cup F_\beta\}$ として作成する．ただし，F は k（k は自然数）個の特徴集合を有し，$F = \{f_1, f_2, \ldots, f_k\}$ である．

(3) TDB_α, TDB_β に対して共通の特徴集合全体から各特徴 $P_\alpha = \{P_{\alpha1}, P_{\alpha2}, \ldots, P_{\alpha k}\}$，$P_\beta = \{P_{\beta1}, P_{\beta2}, \ldots, P_{\alpha k}\}$ の各特徴量 P_x を計算し，特徴量集合 P_α，P_β をそれぞれ作成する．各特徴量 P_x は，下記式 6 を使用して計算する．$|P_x|$ は特徴 P_x の全てのアイテムを含むトランザクションの数を示す．

$$P = \frac{|P_x|}{k} \tag{6}$$

(4) 共通の特徴集合全体が k 個の特徴集合を有する場合，TDB_α と TDB_β 間の距離 D を，下記式 7 を使用して計算する．

$$D = \sum_{i=1}^{k} |P_{\alpha i} - P_{\beta i}| \tag{7}$$

(5) さらに距離 D はアイテム数 k で除する（$D_a = \frac{D}{k}$）ことで 0 から 1 の範囲に収まり，抽出された特徴数に寄らない TDB 間距離を求められる．特徴数の違いを TDB 間の距離として利用するかどうかは分析者の意向による．

3.1 距離の定義との対応

本提案の計算手法が数学的な定義に基づく距離であることを，定義と対応付けて説明する．但し距離という語は擬距離と分けて使われる場合と，擬距離を含む全体を包括して距離と使われる場合がある．本提案の計算手法は正確には擬距離に相当する．距離関数 d が下記の条件 1,2,3 の距離の公理を満たす場合，擬距離と定義される．

(1) 非負性：$d(x, y) \geq 0$

(2) 対称性：$d(x, y) = d(y, x)$

(3) 三角不等式：$d(x, y) + d(y, z) \geq d(x, z)$
さらに，下記の条件 4 を満たす場合，（擬距離と区別される）距離と呼ばれる．

(4) 非退化性：$x = y \rightarrow d(x, y) = 0$ かつ $d(x, y) = 0 \rightarrow x = y$

本提案により計算される距離 D は，絶対値の合計として正の値として算出されるため，上記 1 の要件を満たす．また，共通の頻出アイテム集合を使用し，差の絶対値として計算されることから，入れ替えても同じ値が得られ，上記 2 の要件も満たす．さらに，$x = y = z$ のときに $d(x, y)+d(y, z) = d(x, z)$ であり，上記 1 の非負性を満たすことから，x, y, z が異なる値の場合，$d(x, y)+d(y, z) > d(x, z)$ となり，上記 3 の要件も満たす．従って本提案は少なくとも擬距離となる．

しかしながら，条件 4 においては $d(x, y) = 0$ の場合でも，$x = y$ が成立しないことがある．本提案は TDB から特徴を抽出して比較するため，パラメータによっては特徴が抽出されないことがあり，さらに，違う TDB でもパラメータによって同じ特徴が抽出されることもある．この時に距離は 0 となるが，だからと言って全く同じ TDB であるとは言えない．したがって，距離Dは，正確には擬距離と定義されるが，距離の公理である条件 1,2,3 の要件を満たすことから，距離と表現できる．

3.2 頻出アイテム集合をもとにした距離

距離計算に頻出アイテム集合を用いる手法について説明する．頻出アイテム集合は，同じトランザクションで同時生起する確率の高いアイテムの組み合わせの集合である．頻出と判断すべき同時生起確率の最小値の閾値を最小支持度 $S_{\min}$ といい，これは解析者がパラメータとして入力す

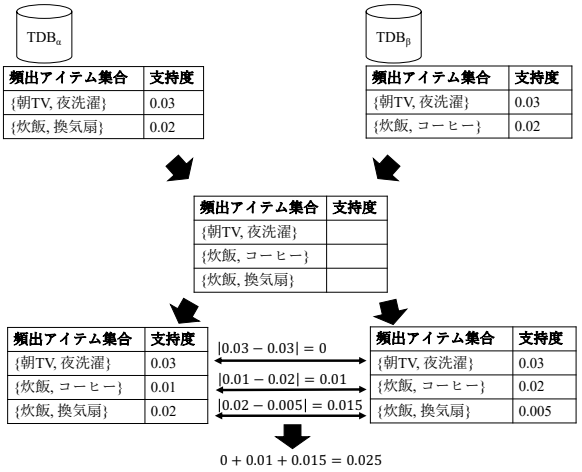


図 3 頻出アイテム集合をもとにした距離
Fig. 3 Distance Based on Frequent Item Sets.

る．支持度 S は式 8 で計算し，最小支持度以上のアイテム集合を頻出アイテム集合として抽出する．

$$S = \frac{\text{集合の全アイテムを含むトランザクションの数}}{\text{全トランザクションの数}} \tag{8}$$

既知の頻出アイテム集合抽出アルゴリズムとして Apriori[4] や FP-growth[5] がある．幅優先探索，深さ優先探索の違いがあるものの，両アルゴリズムでは数学的な lattice 構造を利用して枝刈りをすることで高速に頻出アイテム集合を抽出できる．

この頻出アイテム集合を用いた具体的な距離計算手法を 3 に示す．初めに TDB_α，TDB_β 両方から頻出アイテム集合を抽出した後，頻出アイテム集合全体を作成する．図 3 に示す例では，頻出アイテム集合全体 F が有する個々の頻出アイテム集合 f_k が { 朝 TV, 夜洗濯 } と，{ 炊飯, コーヒー }，{ 炊飯, 換気扇 } である．

各 TDB につき，個々の頻出アイテム集合の支持度を計算する．この時，他の TDB から抽出された頻出アイテム集合に関して支持度を計算することになる．図の例では，{ 炊飯, コーヒー } は TDB_α にとっては頻出アイテム集合ではなかったため，支持度を計算する必要があり，同様に { 炊飯, 換気扇 } は TDB_β にとっては頻出アイテム集合ではなかったためここで支持度を計算する．

最後に，TDB_α と TDB_β で頻出アイテム集合同士の支持度の差の絶対値を計算し，その値を TDB_α と TDB_β との間の距離 D として算出する．図 3 に示す例では，距離が 0.025 と算出される．

3.3 相関ルールをもとにした距離

相関ルールは TDB からの特徴獲得でよく利用される手法であり，頻出アイテム集合におけるアイテム間の相関関係をルールとして抽出するものである．頻出アイテム集合

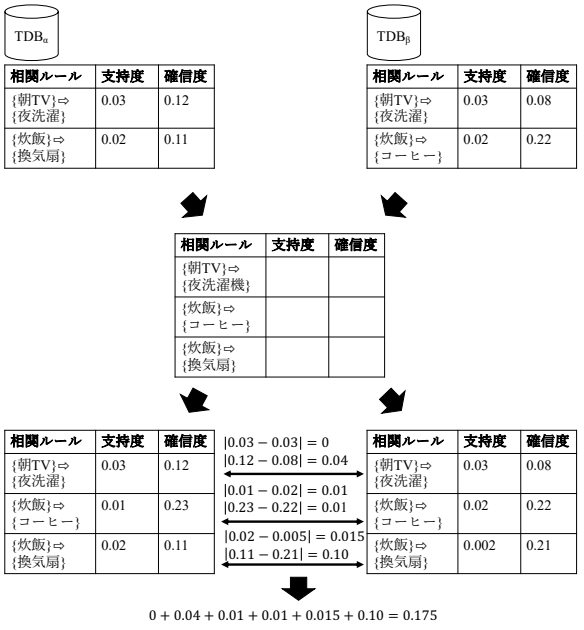


図 4 相関ルールをもとにした距離
Fig. 4 Distance Based on Association Rules.

よりも詳細に DB の特徴を示すと考えられるが，解析者の与えるパラメータチューニングによってルールが抽出できない可能性や，ルール抽出にかかる計算量の増加などの課題もあるため，どちらが優れた手法かは比較できない．本研究では距離計算に相関ルールを用いる手法についても検討した．

相関ルールは頻出アイテム集合のアイテム間の相関を，条件付き確率の発想によって抽出する．具体的には，頻出アイテム集合として { 朝 TV, 夜洗濯 } が抽出された場合，「朝 TV」実施者が「夜洗濯」を実施する確率と，「夜洗濯」実施者が「朝 TV」も実施している確率を計算する．相関ルールの表現としては { 朝 TV} → { 夜洗濯 } とし，左辺のアイテム集合の「朝 TV」実施者が，右辺のアイテム集合の「夜洗濯」実施確率が高いことを示す．この条件付き確率を確信度 C と呼び，式 9 で計算できる．ここで，最小確信度 $C_{\min}$ というパラメータを解析者が入力する．これは相関があると判断すべき条件付き確率の最小値の閾値である．最小確信度 $C_{\min}$ 以上の確信度 C を持つ相関ルールを抽出対象とする．

$$C = \frac{\text{全アイテムを含むトランザクションの数}}{\text{左辺のアイテムを含むトランザクションの数}} \tag{9}$$

この相関ルールを用いた具体的な距離計算手順を図 4 に示す．初めに TDB_α，TDB_β 両方から相関ルールを抽出した後，相関ルール集合全体を作成する．図に示す例では，相関ルール集合全体 F が有する個々の相関ルール f_k が { 朝 TV} → { 夜洗濯 } と，{ 炊飯 } → { コーヒー }，{ 炊飯 } → { 換気扇 } である．

各 TDB につき，個々の相関ルールの支持度と確信度を

Vol.2021-MBL-100 No.6
Vol.2021-UBI-71 No.6
Vol.2021-CDS-32 No.6
Vol.2021-ASD-21 No.6
2021/9/2

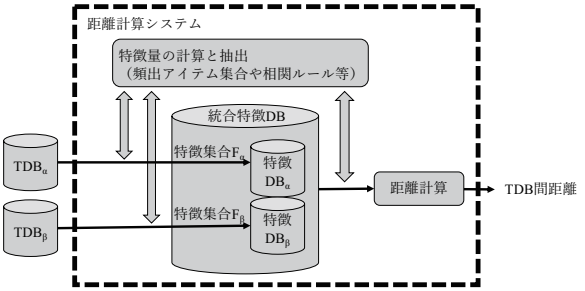


図 5 システム全体
Fig. 5 Entire of the system.

計算する．この時，他の TDB から抽出された相関ルールに関して計算することになる．確信度という特徴量が増えた以外は頻出アイテム集合の時と同様の手続きになる．最後に，TDB_α と TDB_β で相関ルール同士の支持度の差の絶対値と，同様に確信度の差の絶対値を計算し，その値を加算したものを TDB_α と TDB_β との間の距離 D として算出する．図 4 に示す例では，距離が 0.175 と算出される．

4. システム

図 5 は，システムの構成例を示した図である．入力は 2 つの TDB（TDB_α，TDB_β）が必須であり，他にも特徴量の計算と抽出のパラメータが別途必要になることがある．最終的な出力は TDB 間距離であるが，それに利用した特徴量も，統合特徴 DB として確保しておくことで知識発見に役に立つ情報になる。

図 6 は，距離を計算する処理の流れを示したフローチャートである．特徴集合は様々な特徴獲得アルゴリズムを用いることができるが，本研究では頻出アイテム集合にもとづく手法と，相関ルールにもとづく手法を利用した．図中の手順 2 での各 TDB から特徴集合を抽出したときに DB 内にテーブルを作成しておくと，次の手順 3 での和集合は完全外部結合によってシンプルに求める事ができる．

5. 実験

本研究ではダミーデータを作成し，それを提案アルゴリズムで分析することで有効性を検討する．ダミーデータは複数家庭のスマートホーム 1 年分の TDB を想定し，さらに各 TDB でのアイテムの出現頻度を偏らせることによって，アルゴリズムの特徴抽出関数による距離に対する影響を確認する．

アイテム名は記録時間，家電の種類，家電の状態を示す．時間情報は 0 時から 1 時までを 00，23 時から 24 時までを 23 と 2 文字で表現する．家電の種類は炊飯器を R，テレビを T，洗濯機を W，電気ケトルを K と 1 文字で表現する．家電の状態は動作中を R，電源 OFF 動作を D，電源 ON 動作を U，停止中を W，データ取得エラーを N の 1 文字で表現する．ただし，ダミーデータの N は実際のデータ取得エラーを意味していないため，分析やアルゴリズムへの影響は考慮しない．アイテム全体は式 10 のように 4 文字で構成し，アイテムの種類は全部で 480 種類となる．例えば午前 5 時に炊飯器が動作中であれば，05RR というアイテム名になる．

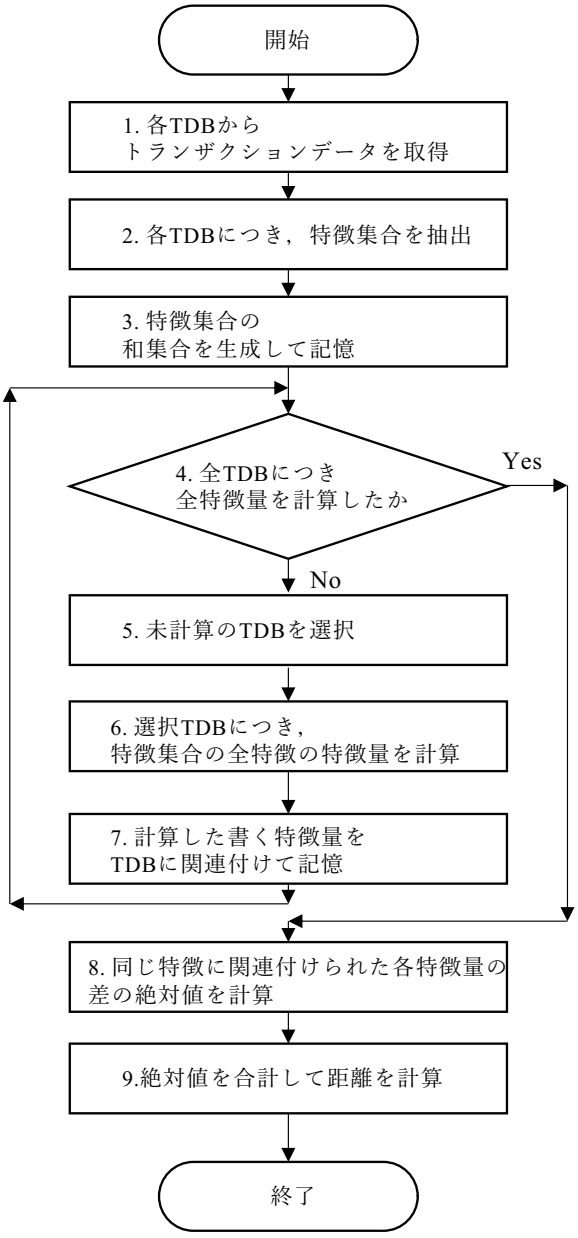


図 6 システム全体の処理手順
Fig. 6 Entire of the flow.

$$[00, 01, \ldots, 23] \times [R, T, W, K] \times [R, D, U, W, N] \quad (10)$$

このアイテムを持つトランザクションを持つ TDB を用意する下記項目は全ての TDB で共通するプロパティである．

- 1 日を 1 トランザクションとする
- TDB 全体は 365 日であり，トランザクション数は 365 とする

- アイテムは 1 時間単位の時間，家電の種類，家電の状態とする

アイテムとして家電の状態に取得エラーである N を用意したので，全軸間において全家電の状態は必ず記録される．従って 1 つのトランザクションは必ず 24 時間分の 4 種類の家電の状態が記録されるため，96 個のアイテムを持つ．そのため各アイテムの出現頻度は状態にだけ依存するので，均等に出現する場合には基本的な支持度は 0.2 となる．これを踏まえて，TDB は 1 から 15 番の 15 種類を用意し，各 TDB はアイテムの出現頻度が異なるように作成した．具体的には次のような種類の TDB を用意した．大分類として Type A，B，C とするが，各 Type の中でさらに小分類がある．

- TDB1 から 5 は全アイテムがランダムで出現する．(Type A)
- TDB6 から 10 はアイテムの出現頻度を偏らせる．(Type B)
  - TDB6,7 は状態 W が多い
    $R : W : U : D : N = 1 : 2 : 1 : 1 : 1$
  - TDB8,9 は状態 U が多い
    $R : W : U : D : N = 1 : 1 : 2 : 1 : 1$
  - TDB10 は状態 D が多い
    $R : W : U : D : N = 1 : 1 : 1 : 2 : 1$
- TDB11 から 15 はアイテムの出現頻度の偏りは抑えつつ，アイテムの共起確率を偏らせる．(Type C)
  - TDB6,7 は 5 の倍数日に W が多い (0.3 程度)，他の日は逆に W が少ない (0.16 程度)
  - TDB8,9 は 5 の倍数日に U が多い (0.3 程度)，他の日は逆に U が少ない (0.16 程度)
  - TDB10 は 5 の倍数日に D が多い (0.3 程度)，他の日は逆に D が少ない (0.16 程度)

この TDB を用いて相関ルール抽出を行った．相関ルール抽出の最小確信度 $C_{\min}$ は 0.30 で固定し，最小支持度 $S_{\min}$ を 0.055, 0.056, 0.058, 0.060, 0.080, 0.100 の 6 パタンで実験した．この値については事前に数回の実験を実施して，経験則的に決めたものである．例えば最小支持度 0.100 以上ではルールが得られない TDB があり，最小支持度 0.054 では相関ルール抽出が終了しない TDB があった．最小確信度も 0.40 では殆どの TDB からルールが得られず，予備実験においておよそ 0.20 から 0.30 程度が良いと思われたが，支持度のほうがルール数の依存度が高く感じられたため，0.30 で固定することにした．

抽出された相関ルール数について Type A，B，C 及び全体での平均ルール数を図 7 に示す．相関ルールの抽出数は Type C が多くなる見込みだったが，全体の支持度は Type A と変わらないため偏り効果は見えなかった．一方で Type B は明らかに抽出されるルール数が増えたため，アイテムの支持度の偏りの効果が見える．

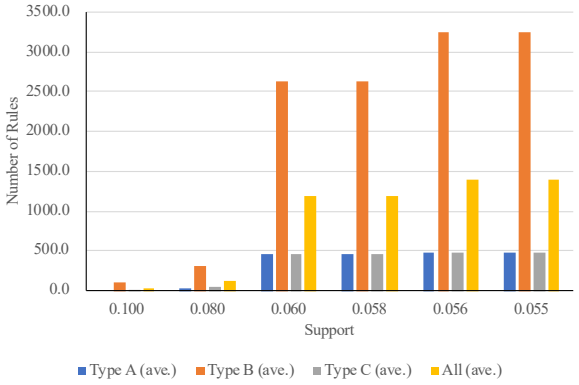


図 7 抽出されたルール数
Fig. 7 Number of extracted rules.

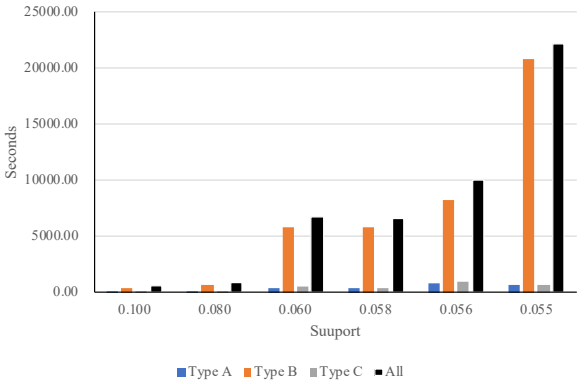


図 8 ルール抽出に要した時間
Fig. 8 Time for extraction.

相関ルール抽出にかかった時間を図 8 に示す．抽出される相関ルール数に依存して処理時間が二次関数的に増加することがわかる．

5.1 距離結果からのアルゴリズム妥当性と有効性の考察

実験結果と，本提案アルゴリズムの妥当性と有効性を考察した．頻出アイテム集合による方法と，相関ルールによる方法に関して結果と考察を順番に説明する．

頻出アイテム集合をもとにした距離計算結果を付録の表 A·1 に示す．今回の頻出アイテム集合には，最もシンプル手法として全アイテムの確率分布を利用している．この距離は抽出した頻出アイテム数で除する方法で計算している．TDB1 と TDB2 の結果は TDB2 と TDB1 の結果と同値であるので黒背景で塗りつぶしている．最右列は最小距離を求めており，それに対応する値を網掛けしている．最小値は各行で算出しているため，列方向は考慮していない．そのため列方向，例えば TDB11 などは列方向の最小値が 5 つあるように見えてしまうことに注意が必要である．

Type A の TDB1 から 5 に関して，全て別の Type の TDB と類似している結果となった．これはこの計算手法が，この Type の TDB の距離を求めるためには機能していない様に見える．Type A の TDB は全アイテムが均等

に出現するため，頻出アイテム集合の支持度では類似性が発見できないことは，提案アルゴリズムにとって妥当な結果と考えられる．

Type B の TDB6 から 10 に関して，同じ Type B の TDB と類似しているという結果になった．特に TDB6 は TDB7 が，TDB8 は TDB9 が，TDB9 からも TDB8 が最小距離という実験結果は望ましい結果であると言える．Type B はアイテムの頻度に特徴をもたせたデータであるので，この計算方法がよく機能している様に見える．

Type C の TDB11 から 15 に関して，Type A の TDB との錯誤が見られる．従って，Type A と Type C はこの計算手法による違いを発見することは難しい事がわかる．一方で，Type A も Type C も，Type B との TDB において最小距離を示すことはなかった．以上の結果から，本提案アルゴリズムによって，頻出アイテム集合という特徴における TDB 間距離を計算できると言える．

相関ルールをもとにした距離計算結果を付録の図 A·2 に示す．この距離は抽出した相関ルールの支持度と確信度の差を単純加算したもので，ルール数で除していない．そのため抽出したルール数の違いもかなり影響している．なお，表の見方は頻出アイテム集合と同様である．

Type A の TDB1 から 5 に関して，Type A 内のみで最小距離を求められた．Type A ではこの計算手法は適している様に見える．

Type B の TDB6 から 10 に関して，Type B 内で最小距離を求めることはできず，全て Type C の TDB とで最小距離が計算された．Type B ではこの計算手法は適していないと考えられる．

Type C の TDB11 から 15 に関して重要な表の分析結果を抽出すると，TDB11 は TDB13 と，TDB12 は TDB13 と，TDB13 は TDB12 と，TDB14 は TDB11 とが最小距離となった．一方，TDB15 は TDB2 とが最小距離となる．Type C 中では TDB15 だけが孤立するように設定したため，他の Type の TDB と最小距離が求まることは予想通りである．一方，TDB11 と 12 がセット，13 と 14 がセットになるようにデータ作成したつもりであるので，これに関して詳細な分析が今後必要になると考えられる．

6. 結論

本論文では類似の行動をする生活者を発見するために，複数の TDB 間の距離計算手法を提案した．生活者の行動による家電の動作状況である DB 内のカテゴリデータをアイテムとして，複数のアイテムが同時共起するトランザクションの集合である TDB を構築する．今回の分析においては，1 日を 1 トランザクションとして暑かった．各生活者の DB が与えられるとき，それかた特徴を抽出してパラメータの差を距離として計算することで，類似の生活者を判断することが可能となる．具体的には，頻出アイテム集合や，相関ルールといった TDB の特徴を利用して，距離を計算することにより，複数の TDB をグルーピングできる．実際にシステムを実装し，アイテム出現確率を偏らせた人為的なダミーデータを複数用意して，距離を計算した．これによって，本提案のアルゴリズムの有効性を実証した．今後の研究として，他の特徴獲得関数による距離の実験，実際のスマートホームのデータを用いた実験，本提案の距離を用いたクラスタリングの実験等が挙げられる．

参考文献

[1] LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning, *nature*, Vol. 521, No. 7553, pp. 436–444 (2015).
[2] McQuitty, L. L.: Hierarchical linkage analysis for the isolation of types, *Educational and Psychological Measurement*, Vol. 20, No. 1, pp. 55–67 (1960).
[3] MacQueen, J. et al.: Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, No. 14, Oakland, CA, USA, pp. 281–297 (1967).
[4] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A. I. et al.: Fast discovery of association rules., *Advances in knowledge discovery and data mining*, Vol. 12, No. 1, pp. 307–328 (1996).
[5] Han, J., Pei, J. and Yin, Y.: Mining frequent patterns without candidate generation, *ACM sigmod record*, Vol. 29, No. 2, pp. 1–12 (2000).
[6] 和世成田，博之北川：トランザクションデータベースに対する高確信度の相関ルールを用いた外れ値検出手法，電子情報通信学会技術研究報告. DE, データ工学, Vol. 107, No. 131, pp. 399–404（オンライン），入手先 ⟨https://ci.nii.ac.jp/naid/110006381653/⟩ (2007).
[7] Deza, M. M. and Deza, E.: Encyclopedia of distances, *Encyclopedia of distances*, Springer, pp. 1–583 (2009).
[8] Norouzi, M., Fleet, D. J. and Salakhutdinov, R. R.: Hamming distance metric learning, *Advances in neural information processing systems*, pp. 1061–1069 (2012).
[9] Levenshtein, V. I. et al.: Binary codes capable of correcting deletions, insertions, and reversals, *Soviet physics doklady*, Vol. 10, No. 8, Soviet Union, pp. 707–710 (1966).

付　　録

– 次ページ –

Vol.2021-MBL-100 No.6
Vol.2021-UBI-71 No.6
Vol.2021-CDS-32 No.6
Vol.2021-ASD-21 No.6
2021/9/2

表 A·1 頻出アイテム集合をもとにした距離計算結果

Table A·1 Distance based on frequent itemsets.

TDB	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Min
1	0.0000	0.0581	0.0394	0.0399	0.0406	0.0436	0.0431	0.0412	0.0433	0.0424	0.0395	0.0391	0.0390	0.0393	0.0392	0.0390
2		0.0000	0.0398	0.0401	0.0406	0.0442	0.0437	0.0423	0.0438	0.0424	0.0397	0.0402	0.0400	0.0400	0.0403	0.0397
3			0.0000	0.0409	0.0418	0.0441	0.0439	0.0418	0.0436	0.0433	0.0399	0.0409	0.0405	0.0408	0.0407	0.0399
4				0.0000	0.0405	0.0436	0.0431	0.0415	0.0432	0.0427	0.0389	0.0399	0.0398	0.0396	0.0400	0.0389
5					0.0000	0.0440	0.0437	0.0410	0.0428	0.0421	0.0387	0.0398	0.0389	0.0393	0.0395	0.0387
6						0.0000	0.0332	0.0594	0.0590	0.0596	0.0624	0.0627	0.0627	0.0623	0.0628	0.0332
7							0.0000	0.0591	0.0589	0.0594	0.0623	0.0623	0.0625	0.0620	0.0626	0.0589
8								0.0000	0.0310	0.0577	0.0604	0.0609	0.0605	0.0603	0.0609	0.0310
9									0.0000	0.0587	0.0617	0.0624	0.0620	0.0618	0.0623	0.0587
10										0.0000	0.0615	0.0623	0.0618	0.0617	0.0616	0.0615
11											0.0000	0.0406	0.0409	0.0406	0.0408	0.0406
12												0.0000	0.0405	0.0397	0.0399	0.0397
13													0.0000	0.0398	0.0406	0.0398
14														0.0000	0.0406	0.0406

表 A·2 相関ルールをもとにした距離計算結果

Table A·2 Distance based on association rules.

TDB	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Min
1	0.00	195.61	181.98	183.66	196.80	708.80	843.01	579.87	778.24	671.90	190.99	188.09	192.55	189.30	194.41	181.98
2		0.00	182.52	178.69	182.21	705.22	808.63	570.28	820.43	669.86	197.59	192.01	185.81	192.55	186.23	178.69
3			0.00	185.55	183.57	709.17	801.22	579.68	779.01	677.52	199.17	189.17	193.93	188.96	186.96	183.57
4				0.00	192.21	716.53	825.81	579.79	702.39	746.23	195.83	194.59	193.47	193.58	198.69	192.21
5					0.00	706.62	796.60	577.35	848.17	669.95	195.13	185.83	191.93	186.70	194.87	185.83
6						0.00	2536.02	2299.87	2507.28	2453.50	1944.35	1957.62	1929.80	1935.43	1939.76	1929.80
7							0.00	2171.07	2369.25	2350.57	1807.65	1828.32	1799.97	1808.12	1800.28	1799.97
8								0.00	1991.20	1833.12	1296.48	1284.38	1316.25	1328.14	1282.37	1282.37
9									0.00	2353.54	1803.52	1794.92	1836.91	1846.62	1796.23	1794.92
10										0.00	1704.88	1680.61	1690.98	1681.77	1739.50	1680.61
11											0.00	192.74	178.92	185.51	190.76	178.92
12												0.00	188.20	191.70	191.08	188.20
13													0.00	197.92	190.54	190.54
14														0.00	193.68	193.68