情報処理学会論文誌　データベース　Vol.14　No.1　18–29 (Jan. 2021)

事例・実践論文

不動産取引データベースの網羅性向上を目的とした不動産募集広告情報のレコード同定

馬場 弘樹[1,a)]　関口 知子[1,2,b)]　清田 陽司[1,2,c)]　清水 千弘[1,d)]

受付日 2020年6月10日, 採録日 2020年9月30日

概要：不動産取引の網羅的な捕捉にあたって有力な情報源となる不動産募集広告情報は，多数の不動産仲介業者によって作成されているため，情報の重複や精度など，利用にあたって解決すべき課題がある．特に，都市部の賃貸市場において大多数を占める集合住宅物件（マンション・アパートなど）の情報には多数の重複がみられ，同一の部屋（住戸）を集約する作業が必須である．本論文では，同一物件である可能性の高い住戸情報を集約するタスクを，データ工学におけるレコード同定問題の一種として定義し，文献・個人・商品などの集約タスクとの性質の違いを示したうえで，クラスタリングなどの既知のデータ処理手法を適用することで，実用的な精度が達成できるかどうかを検証した結果を報告する．

キーワード：不動産物件情報，レコード同定，クラスタリング，棟寄せ，戸寄せ

Record Linkage of Real Estate Advertising Information to Improve Comprehensiveness of a Real Estate Transaction Database

Hiroki Baba[1,a)]　Tomoko Sekiguchi[1,2,b)]　Yoji Kiyota[1,2,c)]　Chihiro Shimizu[1,d)]

Received: June 10, 2020, Accepted: September 30, 2020

Abstract: Real estate information database, an influential data source for comprehensive understanding of real estate transactions, has some problems, since the database is created by multiple real estate intermediary agents. Particularly, we confirm that there are substantial duplication in condominiums and apartments, and thus, it is necessary to integrate the duplicate records together. We regard the task as one of record linkage problems, and develop the model integrating the high likelihood of dwellings with the application to existing data handling techniques such as hierarchical clustering. We then validate whether the integrated records by the proposed method achieve practical recall and precision.

Keywords: real estate information, record linkage, clustering, building linkage, dwelling linkage

1. はじめに

不動産取引市場（売買・賃貸）は不透明であり，その取引において高いコストが発生するといったことはかねてから指摘されてきた．家を売りたい，買いたい，または貸したい，借りたいといった家計は，不動産情報が不足することで，高い探索費用が発生しているのである [1]．そのような費用が発生する原因としては，正しい市場価格が把握できないという問題 [2] のほかに，情報流通において表記あるいは数値の揺れや抜け漏れ，誤入力といった，さまざまなエラーが混入する要因が，不動産市場の内部に存在していることがあげられる．つまり，不動産市場を取り巻く制度や技術的な制約からもたらされる情報の精度・確度が低いといった問題のほかに，不動産市場のプレーヤによってもたらされる人為的な問題に起因するエラーが存在しているのである．そのようなエラーは取引コストを高めるだけでなく，不動産バブルを生み出す原因にもなることで，社

1　東京大学空間情報科学研究センター
Center for Spatial Information Science, The University of Tokyo, Kashiwa, Chiba 277–8686, Japan
2　株式会社 LIFULL
LIFULL Co., Ltd., Chiyoda, Tokyo 102–0083, Japan
a)　hbaba@csis.u-tokyo.ac.jp
b)　sekiguchitomoko@lifull.com
c)　kiyotayoji@lifull.com
d)　cshimizu@csis.u-tokyo.ac.jp

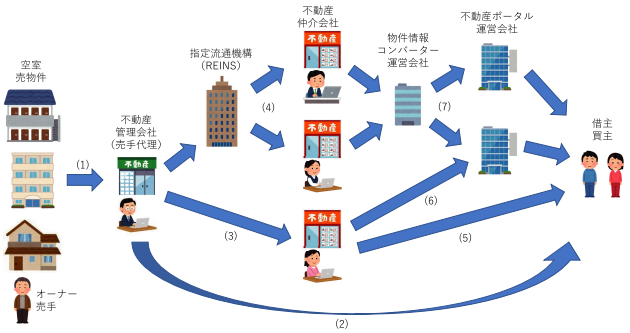

図 1　不動産募集広告情報流通の仕組み

Fig. 1　Overview of propagation of real estate information.

会全体の厚生水準を低下させるだけでなく，しばしば社会全体に甚大な混乱をもたらしてきた [3].

このようなエラーは，不動産市場のプレーヤが，オンライン上での不動産募集広告が一般的になるにつれて，ポータルサイト上で自分が出した広告に対して高い反響を獲得したいという動機が高まることによっても引き起こされる．ゆえに，同じ不動産募集広告であったとしても，異なる不動産物件の広告であるようにポータルサイトに掲載されることが発生している．

このような問題を正しく理解するために，不動産募集広告情報が生産・流通する一般的な仕組みを**図 1** に示す．賃貸物件において空室が生じると，物件のオーナは，契約している管理不動産会社を通じて入居者募集を依頼する (1)[*1]．管理不動産会社は，直接借主（買主）を募集する場合 (2) もあるが，多くの場合は，より広く募集を行うため，提携している不動産仲介会社に依頼 (3)[*2] したり，REINS[*3] というシステムを通じてより多くの仲介不動産会社に募集を依頼 (4) したりする．不動産仲介会社は，自ら広告を出して借主（買主）を募集 (5) することもあるが，多数の物件探しユーザを抱える不動産ポータルに掲載して募集することも多い (6)．また，複数の不動産ポータルに掲載する場合には，物件情報コンバータなどのサービスが利用される (7) こともある．

このような過程のなかで，1 つの不動産募集広告であったとしても，異なる不動産の広告であるように扱われてしまうことが起こりうる．不動産仲介業者は，買い手，借り手を見つけて契約が成立して初めて報酬を得ることができる．1 つの不動産をめぐって複数の業者が広告を出すために，それが集約されてしまうと，買い手・借り手からの問合せが自分に入る確率が低下してしまうために，違う広告として扱われることへの動機が存在するためである．

一方で，消費者にとっては，1 つの不動産物件の広告であったとしても，複数の情報としてポータルサイトなどで掲載されると，探索費用が上昇するだけでなく，不動産市場に対する不信感を増幅させてしまうという問題が発生する．その意味で，不動産市場で流通している情報の品質，とりわけ同一の情報の集約，いわゆる「名寄せ」をしていくという行為は，きわめて重要となるのである．

名寄せには，マンション・アパートの単位で特定していくという「棟寄せ」とそのマンション・アパートの内部に存在する住戸・部屋単位，つまり「戸寄せ」といった階層に分類される．

情報の品質および棟寄せ・戸寄せ処理に関しては以下にあげるような課題がある．

(a) 標準情報フォーマットの不在　業界団体による自主規制ルール[*4]により，不動産募集広告の表示項目についてはある程度の標準化がなされているものの，コンピュータ処理が可能で広く普及している標準情報フォーマットが現時点で存在しない[*5]．

(b) 作業ミスや判断の揺れによるエラーの可能性　標準情報フォーマットが存在しないことから，同一の募集物件に対して，募集に関わるそれぞれの不動産会社が人手での募集広告情報の作成を強いられる場合も多く，作業ミスや判断の揺れにともなうエラー（記入漏れ，誤字・脱字，表記揺れ，数値の丸めなど）が発生しやすい．

(c) 集約処理の回避によるエラーの可能性　同一の募集物件に対して複数の不動産会社が募集広告情報を作成するということは，ユーザが不動産ポータルで物件を検索する際に，同一の物件が多数表示されてしまうという不便につながるため，不動産ポータル側では集約処理を行う[*6]ことでユーザビリティを向上させるなどの取り組みを進めている．一方で，自社の募集情報が他社のものに集約されてしまう可能性のある不動産会社にとっては，占有面積や住所などの表示項目をわずかに変化させたりすることで，集約処理を回避し，自社の募集情報をより目立つ形でポータルに表示させるという動機が存在する．

上記にあげた課題を解決するためには，ミス，判断の揺れや集約処理の回避によるエラーの存在を前提とした同一物件の集約手法が必要とされる．「同一の実体を指す複数の情報を集約する」というタスクは，一般に「名寄せ」

[*1] 物件売却の場合も同様に媒介契約した不動産会社に買主の募集を依頼する．

[*2] 募集条件，立地，建物概要，間取り図などの物件情報を 1 枚のシートにまとめた通称「マイソク」とよばれる資料によって行われることが多い．

[*3] Real Estate Information Network System（不動産流通標準情報システム）の略称．国土交通省が宅地建物取引業法に基づき指定する組織である指定流通機構が運営する．日本国内では 4 つの指定流通機構が存在し，それぞれの会員不動産会社のみがアクセスできるデータベースシステムとして運営されている．

[*4] 不動産の表示に関する公正競争規約
http://www.rftc.jp/koseikyosokiyaku/

[*5] XML などで業界標準フォーマットを定義しようという動きはある [4] が，広く普及しているとはいえない．

[*6] 物件ごとにユーザにとっても最も価値が高いと考えられる募集広告情報を優先的に表示するなどの処理が一般に行われる．

とよばれるが，データ管理の汎用問題であるレコード同定（record linkage）[5] とも見なすことができる．古くは図書館における書誌目録作成の過程で生じる「著者名」や「書名」の集約のため行われてきた典拠管理から，近年では EC サイトにおける同一商品情報の集約まで，さまざまなタスクを対象として手法が開発され，適用されてきた．

それでは，不動産募集広告情報を対象としたレコード同定タスク（以下，「不動産レコード同定」という）には，どのような性質があるのだろうか？ 詳しい議論は次節に述べるが，不動産レコード同定は，対象の「特定性」「階層性」「市場性」「情報非対称性」の 4 点において，他のレコード同定タスクと大きな違いがある．不動産レコード同定の実応用にあたっては，これらの性質をふまえた手法の選択が必要とされるが，不動産における棟寄せ・戸寄せなどのタスクの手法や精度を，このような観点から論じた先行研究は，著者らの知る限り存在しない．

本研究では，不動産募集広告情報から，クラスタリングなどの既知のデータ処理方法を組み合わせることで，物件情報のレコード同定を自動処理できる手法を提案し，実用的な精度が達成できるか検証した．手法構築の際には，不動産レコード同定の特質である情報の非対称性や階層性に注意を払い，物件間の属性の誤差を一定程度許容するとともに，棟単位でのマッチングを意識したうえでの戸寄せを行い，精度の向上を図った．

不動産情報の集約には，主に即時的な高品質・高精度かつ確度の高い情報の提供，集約処理の回避にともなうエラーを抑制することに意義があるといえる．売り手・買い手といった消費者は，市場の温度感（価格や家賃）を捕捉したいために，今でも公示地価または路線価などの政府が公表する価格情報を参考にしていることが多いが，それらは測定されてから公表されるまでに短くとも数カ月のタイムラグが生じる．そのようなタイムラグはしばしば不要なバブルの要因となるため，正確かつリアルタイムな不動産情報が重要であるということは，リーマンショックの反省を受けて IMF や国際決済銀行といった国際機関からも指摘されている [6]．このようなマクロ的な問題に加えて，効率的な不動産情報の提供，ストックの把握といったミクロな問題は，直接的な市場の信頼度に影響をもたらす．これらのマクロ・ミクロの両面から，実用的な水準でレコード同定を行うことは高い社会的な意義を持つといえる．

2. 本研究の位置づけ

前述のとおり，不動産における棟寄せ・戸寄せなどのタスクは，レコード同定の一種と位置づけることができる．

図書館などの文献目録作成においては，著者や主題などの各種の概念に対して，一貫した識別子を付与するための典拠管理（authority control）[7] が，古くから行われてきた．たとえば，「ウィリアム・シェイクスピア」「Shakespeare, William」「沙士比阿」などの著者名を同一の著者を指すものとして結び付けて管理することで，文献検索などの利便性を確保してきた．典拠管理は，カタロガーとよばれる専門職によって行われる高度な判断を要する作業である．このように，もともとはレコード同定に属する処理は，専門職の人手に委ねられる作業であった．

第二次世界大戦期におけるコンピュータの実用化と期を同じくして，図書目録以外を対象としたレコード同定にも目が向けられた．Dunn [5] は，米国の公衆衛生学分野のジャーナルにて，文献の著者に限らず，あらゆる個人を対象として出生から死亡までの一貫した記録を管理するための方法論を提案している．当時から，すでに同姓同名，表記揺れ，データの誤りなど，レコード同定にあたって解決すべきエラーの問題が指摘されていた．その後，1950 年代から 1960 年代にかけて，Newcombe ら [8]，Fellegi ら [9] が，エラーの発生過程を統計的にモデル化することにより，レコード同定を数学的な問題として定義した．この頃は，社会的・医学的な統計調査などが主な応用であり，訓練された専門家によって作成されたデータを対象としている．

1990 年代後半からのインターネットおよび Web の普及，それにともなう Web 上のさまざまなビジネスの発達は，専門家ではない多数の人々（たとえば EC サイトの出品者など）によって作成，共有されるデータ量の激増につながった．こうしたデータには，専門家によって作成される文献目録，調査票などと異なり，情報の重複やエラーが数多く存在する．Web 上の情報サービスでは，非専門家によって作成された種々のデータのレコード同定を行い，検索などの利便性を高めることへのニーズが高まった．また，機械学習アルゴリズムの発達，訓練用データセットの整備にともない，機械学習を用いたレコード同定の手法 [10] とともに，人名検索サイト，書籍販売サイト，家電販売サイト，オークションサイトなど，さまざまなサービスにおけるレコード同定も研究されるようになった．Wikipedia を用いた大規模な固有表現の曖昧性解消 [11]，Web ページ上の人名の名寄せ [12]，書籍販売サイトにおける著者名の名寄せ [13] など，さまざまなタスクを対象としたレコード同定が研究されている．

本論文で対象とする不動産レコード同定の性質を，他の対象におけるレコード同定と比較したものを**表 1** に示す．不動産レコード同定を特徴付けるのは，以下に示す 4 点である．

a) 特定性（specificity）：まったく同一の実体が 2 つとして存在せず，実体の特定が容易である
b) 階層性（hierarchy）：上位クラスの実体と下位クラスの実体の間に包含関係がある
c) 市場性（marketability）：実体が市場で取引され，需要と供給による価格メカニズムがはたらく
d) 情報非対称性（information asymmetry）：市場におい

表 1　レコード同定タスクの比較
Table 1　Comparative tasks over record linkage.

対象	具体例	a) 特定性	b) 階層性	c) 市場性	d) 情報非対称性	レコード同定の課題	応用サービス例
人物	著者 個人 Web ページ	○	-	-	-	個人情報保護強化の流れ 名称の曖昧性（同名，別姓，ニックネーム，異体字，改名など）	図書館 国民保険情報管理 文献検索サイト 人物検索サイト
組織	法人，団体	○	△	△	○	名称の曖昧性	企業信用情報サイト
著作物	本，論文	-	△	△	-	版，翻訳，媒体の多様性など 属性の欠損値，誤り	図書館 文献検索サイト 書籍販売サイト
工業製品（新品）	家電製品	-	△	○	-	製品改良による仕様変更など	EC サイト
工業製品（中古品）	中古車	△	△	○	○	中古市場における状態の多様性	オークションサイト フリーマーケットサイト
サービス	労働力	△	-	△	○	品質やスキルの多様性	求人サイト クラウドソーシングサイト
不動産	建物，土地	○	○	○	○	属性の欠損値，誤り 集約処理の回避によるエラー	不動産ポータルサイト

て，売手と買手の間で同一の情報の共有ができておらず，人為的な問題に起因する情報のエラーが起こりやすい

上記の 4 点のうち，a) および b) はレコード同定を容易にする要因として作用する．a) については，本などの著作物[*7]，車や家電製品などの工業製品[*8]などと異なり，不動産物件は「同一の実体か否か」という定義における曖昧性が存在せず，住所・地番などによって一意に特定できるため，精度などの評価も比較的容易である[*9]．また，b) については不動産には部屋（戸）→ 建物（棟）→ 土地（筆）という明確な階層関係があるため，いずれかのレコード同定（例：戸寄せ）の結果が他のレコード同定（例：棟寄せ）にも手がかりとして利用できる．組織や著作物，工業製品などにもある程度の階層性はあるものの，カテゴリの曖昧性などが存在するため，手がかりとして利用できない場合も多い[*10]．

c) は，要求される情報集約の精度に影響を与える．市場性のない（もしくは乏しい）対象[*11]への entity linkage タスク（著作物，個人など）では，そもそも高い精度で集約できていなければ，実用上のメリットは得られない．しかし，不動産のように価格メカニズムがはたらく対象においては，たとえ集約の精度がそれほど高くなくとも，周辺の取引事例情報を通じて需要・供給の動向を把握できること，参考価格や不動産価格指数の算出など，実用上のメリットを得られるアプリケーションは多い．一方で，不動産ポータルサイトなどにおける物件情報の集約においては，顧客である不動産会社の利害に直結することから，高い精度が求められる．

d) は逆にレコード同定を困難にする要因として作用する．情報を意図的に操作することで利益を得られる[*12]という構造が存在するため，ある程度の人為的な要因によるエラーの存在を前提としなければならない[*13]．

本研究では，クラスタリングアルゴリズムの一種である階層的クラスタリングを適用することで，不動産レコード同定を行う手法を提案する．不動産情報処理分野では，Harrison らによる先駆的な研究 [14] をはじめとして，価格推定にク

*7　著作そのものと，版・刷による内容の違い，翻訳版の存在，媒体（ハードカバー，文庫，電子版）などの多様な形態のいずれを指しているか特定することが困難である．
*8　大量生産される工業製品，特に新品として流通している家電製品などでは，市場でやりとりされる情報は型番などであり，実体を特定する製造番号などはやりとりされない．製品改良による細部の仕様変更などは，型番などに反映されないことも多い．中古品についても，製造番号などが明らかにされることはほとんどなく，特定は必ずしも容易ではない．
*9　表 1 では，工業製品（中古品），サービスについても原理的には特定が可能であるが，特定に用いることができる手がかり情報（製造番号，組織内における ID など）は一般に流通せず，実務上特定が困難であることから△としている．
*10　表 1 では，カテゴリの曖昧性などにより，手がかりとしての利用が困難であることをもって△としている
*11　表 1 では，同等の財での代替が可能かどうかという観点で分類している．組織や著作物，サービスは，需要と供給がある程度価格に反映される一方で，代替が著しく困難な場合も多いことから，△としている．
*12　前述の不動産ポータルにおける集約の回避のほか，いわゆる「おとり物件」の問題などもある．すでに入居者が決まっている物件について「問合せがたくさん入るから」という理由で意図的に募集広告の掲載を削除しなかったとしても，「たったいま契約が決まってしまいました」といわれれば，一般ユーザがおとり物件であることを見抜くのは難しい．
*13　この問題は，中古の工業製品，サービスの取引，企業間の M&A などにも共通する．

ラスタリングを用いる研究がいくつも行われている [15]．しかし，クラスタリングをレコード同定に適用し，精度などを評価した先行研究は，著者らの知る限り存在しない．

特定性や市場性が存在しない，もしくは乏しい対象（人物，組織，著作物，工業製品，サービスなど）については，クラスタリングの適用が有用かどうかは，アプリケーションに依存する．Ono らの研究 [12] のように，情報検索を目的とする場合には，精度や再現率が非常に高くなくても有用であるが，高い精度・再現率が求められるようなアプリケーションでは，適用が実用上難しい．よって，レコード同定については，EAN コード[*14]や Amazon の ASIN コード[*15]のように，中央集権的な組織によって割り当てられる識別子や，クラウドソーシングなどの人間による判断が利用されることが，実用上ほとんどである．

門らの研究 [16] は，集合住宅の棟に関するレコード同定を扱い，棟の名称の文字列，住所コード，階数，築年月の情報を用いて 2 つのレコードの一致・不一致を 4 層ニューラルネットを用いて判定させる手法を提案している．扱っている問題は本研究と類似しているものの，本研究は，不動産レコード同定のタスクの性質をふまえたより一般的な位置づけの議論を示している点で門らの研究と異なる．また，提案手法および評価の観点では，棟より細かい住戸単位の募集情報についての戸寄せを対象としている点，棟の名称が欠損値であるケースも多いことをふまえ，棟の名称を利用しない場合の精度評価を行っている点で異なる．

3. データ

本研究では，2008 年 1 月 1 日から 2019 年 11 月 30 日において LIFULL HOME'S に掲載された居住用の賃貸マンション・アパートの物件データを用いた．当該データは不動産仲介業者から依頼のあった物件をウェブサイト上に掲載したものであり，前述したように複数の業者が同一の物件を掲載している．たとえば，世田谷区を例にあげると，上記期間の居住用賃貸マンション・アパートのレコード数は約 523.0 万件であるが，実際には物件データの重複が発生しているため市場に出現している物件数は上記のレコード数よりも小さい．

物件データの特徴としては，以下の点があげられる．当物件データは物件の成約などにより掲載終了するまでに，募集条件の変更などを反映するため 2 週間未満の頻度で更新される．したがって，掲載開始時と掲載終了時で家賃などの情報が異なることがある．この問題に対処するため，掲載終了時の情報を基準値として抽出した．さらに，当物件データの一部には各マンション・アパートに対してあらかじめユニークに割り当てた棟 ID[*16]という識別番号が存在する．しかし，棟 ID は信頼度が高い一方で，たとえば世田谷区では棟 ID が入力されているレコードの割合は 73.4%にとどまり，棟 ID だけでマンション・アパート棟を特定することはできない．

分析に用いた属性は住所，物件名，棟 ID，築年月，建物階数，専有面積，間取り，部屋階数，部屋番号である．棟単位の属性は，棟 ID に加えて住所，物件名，築年月，建物階数を用いた．なお，本研究では戸寄せの手法提案を目的としており，利用データは棟と戸の属性を明示的に階層化させている訳ではない．それでも，棟単位でのマッチングを意識したうえでのモデル構築が重要であるという考えから，信頼性の高い棟 ID の重みを変化させるように次章で考える．

戸単位での属性は専有面積，間取り，部屋階数，部屋番号であるが，前 3 者が正確に入力されていたと仮定しても，類似する規格の住戸が並ぶアパートなどでは，特定が困難である．部屋番号は正確に住戸を特定でき，世田谷を例にとると 98.9%と高い入力率であるが，情報の正確性について担保されている訳ではない．たとえば，「1301 号室」は「1301」，「1301 号室」，「1-301 号室」などさまざまな表記法による揺れが存在する．したがって，部屋番号をベースにしつつ，専有面積，間取り，部屋階数を参照しながら住戸を特定するような技術の考案が必要となる．

続いて，データの前処理について述べる．住所は大きく都道府県，市区町村，町名，街区番号，住居番号に分かれる．たとえば，東京都千代田区千代田 1 丁目 1-1 は，「東京都」，「千代田区」，「千代田 1 丁目」，「1」，「1」と分けられる．本研究では東京大学空間情報科学研究センターが開発・提供しているパッケージである DAMS（Distributed Address Matching System）[*17]を利用して正規化を行った．DAMS は街区番号レベルまでを特定できるため，街区番号までの住所を利用して正規化した．

物件名は英数字の半角全角やひらがなカタカナといったさまざまな表記揺れを含む．したがって，前処理段階で全角の混在する文字列を半角に統一し，大文字を含む英数字を小文字に正規化した．なお，英単語の表現（たとえば，「mansion」と「マンション」）などの言語表記の違いもレコード同定に一定程度の影響を与えると考えられるが，本研究では扱っていない．

部屋番号も物件名と同様にさまざまな表記揺れを含むため，半角，小文字への正規化に加え，以下のとおり具体的に処理を行った．まず，文字「号」「室」「番」「館」を除去

[*14] 商品識別およびバーコードの国際的規格の 1 つ．日本では JAN コードとよばれる．

[*15] https://www.amazon.co.jp/gp/seller/asin-upc-isbn-info.html

[*16] 住宅地図会社が保有するデータベースと API 経由で突合し，住所や物件名が一致する場合に住宅地図上の建物に割り振られた ID を取得したもの．

[*17] 位置参照技術を用いたツールとユーティリティ http://newspat.csis.u-tokyo.ac.jp/geocode/modules/dams/index.php?content_id=1

し，「階」を「F」に変換した．さらに，数字の後に続く最後の文字列を削除した．これにより，たとえば「2 階 2 0 3 号室 A」は「2F203」に変換される．

4. 提案手法

前処理したデータを用いて，不動産募集広告情報についてレコード同定を行うが，その提案にあたっていくつかの課題が存在する．第 1 に，物件名の表記揺れが大きいため，同一物件の相違度を測るための変換を行う必要がある．第 2 に，各物件属性について，その信頼性が異なるため，物件間の情報の相違度を確定する必要がある．第 3 に，各物件属性間での信頼性の重み付けを設定する必要がある．

そこで，本研究では以下のとおり課題に対処した．第 1 に，本研究ではバイグラムによる文字列の分解を行い，1 からコサイン類似度を差し引くように定義されたコサイン相違度により物件名間の距離を測定した．これにより，たとえば「マンション」と「マンシオン」の違いは前処理による完全一致が困難であるが，提案手法による距離では両者間である程度近い距離を算出できる．第 2 に，各属性相違度をを 3 水準で区分し，定量的な評価ができるようにした．ここでは，属性相違度 A をほぼ同一属性であると見なし，属性相違度 B を同一とはいい切れないものの，類似している物件，それ以外を属性相違度 C として定義した．第 3 に，属性相違度によってペナルティをつける度合いを変化させることで対応した．特に，信頼性の高い棟 ID と，内覧の際に誤った情報を提供しづらい部屋階数，住戸を唯一特定できる部屋番号は重みを大きく設定した．

以上をふまえた戸寄せ技術の手法は，集約処理の回避や作業ミスなどによるエラーの影響を考慮できるような物件間相違度を距離で表すことができると考えられる．本研究では，**図 2** の流れで不動産情報の集約を行った．

以下に，提案手法の流れについて具体的に述べる．

はじめに，物件名を 2 個の文字でそれぞれ分割するバイグラム処理を行った．手法としては，文字数 a と文字数 b でバイグラムを作成し，c 個の文字列重複がある場合，$1 \times (a+b-c-2)$ の文字列ベクトルを用いて，コサイン相違度によって 2 物件間の距離を算出した．コサイン相違度は物件名 A，B の文字列ベクトルをそれぞれ $\vec{u_A}$，$\vec{u_B}$ としたとき，

$$1 - \cos(\vec{u_A}, \vec{u_B}) = 1 - \frac{\vec{u_A} \cdot \vec{u_B}}{|\vec{u_A}||\vec{u_B}|}$$

と表せる．たとえば，「麹町アパート」は「麹町」「町ア」「アパ」「パー」「ート」にそれぞれ分割される．これを「麹町ハウジング」と比較する場合，「町ハ」「ハウ」「ウジ」「ジン」「ング」を加え，該当する文字列を 1，それ以外を 0 とする 10 次元のベクトルによって比較する．

続いて，2 物件間について各属性相違度を測定した（**表 2**）．住所について，番地まで完全一致のものを属性相違度 A，番地の不一致があるものを属性相違度 B とした．物件名では，コサイン相違度によって測定した距離が 0，すなわち完全一致の場合を A，0 以上 0.3 未満の場合を B，それ以外を C と評価した．棟 ID は誤差の存在する可能性が低いため，完全一致の場合のみ A，それ以外を C とした．築年数は 1 カ月の誤差を A，1 年間の誤差を B, それ以外を C とした．建物階数および部屋階数は完全一致の場合を A，1 階分の誤差がある場合を B，それ以外を C とした．専有面積は比較対象面積の差分が 10%未満であれば A，20%未満であれば B, それ以外を C とした．間取りは完全一致の

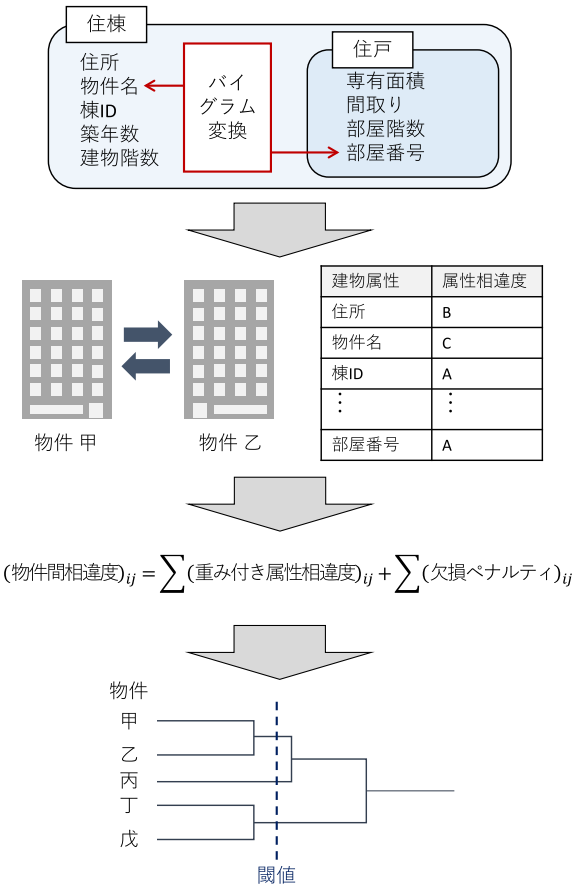

図 2　提案手法のフロー
Fig. 2　Procedure for the proposed method.

表 2　各属性相違度
Table 2　Degree of difference in real estate attributes.

属性		属性相違度 A	属性相違度 B
棟単位	住所	完全一致	番地の不一致
	物件名	0	(0, 0.3]
	棟 ID	完全一致	—
	築年月	<1 カ月	1 カ月 ≤, <12 カ月
	建物階数	0 階	1 階
戸単位	専有面積	差分/面積 <0.1	0.1 ≤ 差分/面積 <0.2
	間取り	完全一致	部屋数は異なるが総部屋数が一致
	部屋階数	0 階	1 階
	部屋番号	[0, 0.3]	(0.3, 0.6]

場合が A で，部屋数は異なるがリビングなども一部屋と考えた場合の総部屋数が一致している場合に B，それ以外を C とした．たとえば，3DK と 2LDK は部屋数がそれぞれ 3 と 2 であり，間取り種類も異なるが，総部屋数でみると 5 部屋と解釈できるため属性相違度 B となる．最後に，部屋番号は物件名と同様にコサイン相違度を算出したが，物件名よりも文字数が少なく，相違度が高くなる傾向にあるため，0 以上 0.3 未満の場合を A，0.3 以上 0.6 未満の場合を B，それ以外を C とした．

以上をふまえ，各属性相違度の総和として物件間相違度を以下のように定義した．

$$S_{ij} = \sum_{k \in K} s_{k,ij} + \sum_{l \in L} (100 * s_{l,ij} + p_{l,ij})$$

ただし，S_{ij} は i 番目と j 番目物件の物件間相違度，K ＝ {住所, 物件名, 築年月, 建物階数, 専有面積, 間取り}，$s_{k,ij}$ は k 番目属性について属性相違度 A で 0，B で 1，C で 2 を返す関数，L = {棟 ID, 部屋階数, 部屋番号}，$p_{l,ij}$ は l 番目属性の欠損ペナルティを表す．たとえば，信頼性の高い棟 ID の場合，属性相違度の重みを 100 とした．このとき，両者とも棟 ID があり，一致していれば $100 * 0 = 0$ を，一致しない場合に $100 * 2 = 200$ を物件間相違度に加算した．一方にのみ棟 ID が存在する場合，あるいはどちらも棟 ID が存在しない場合，欠損ペナルティを 1 とした．また，部屋番号は住戸を唯一ユニークに特定でき，部屋階数も住戸の特定に大きく寄与する属性のため，棟 ID と同様の物件間相違度加算を行った．すなわち，両者とも部屋番号（部屋階数）があり，属性相違度 A の場合に $100 * 0 = 0$，属性相違度 B の場合に $100 * 1 = 100$，属性相違度 C の場合に $100 * 2 = 200$ を加算した．一方のみ，あるいはどちらも部屋番号（部屋階数）が存在しない場合には欠損ペナルティを 1 加算した．このように各属性相違度に重みづけとペナルティを付与することにより，物件間相違度が小さいほど両物件が類似しているといえる．

最後に，算出した物件間相違度に基づき，階層的クラスタリングを行うことで類似物件を集約した．階層的クラスタリングは最短距離法を用いて求めており，物件間相違度を下三角距離行列に拡張して，物件間の距離が小さいペアから逐次クラスタを作成していった．

なお，評価については検証対象地を設定し，目視により作成した正解データと比較した．レコード同定で重要となるのが同一物件の総数のうちどの程度同一物件を捕捉できているかを示す再現率（$\frac{TP}{FN+TP}$），レコード同定対象とした物件のうちどの程度正解しているかを示す適合率（$\frac{TP}{FP+TP}$）である．ただし，TP は真陽性，FN は偽陰性，FP は偽陽性の度数をそれぞれ表す．両者はトレードオフの関係にあり，再現率を上げるほど不一致の物件が混入するため，適合率が下がってしまう．したがって，相違度の閾値を変化させながら，両者を比較衡量してクラスタリングを行った．

5. 提案手法の精度検証

提案した手法について精度を確認するため，東京都世田谷区，足立区，千葉県船橋市を対象として精度検証を行った．検証に際して正解データを作成する必要があるが，上記対象地からそれぞれ 999 件，746 件，729 件を抽出し，目視により正解棟番号，正解部屋番号を付与した．

正解データのためのレコード抽出について，レコードがすべて同じ棟である場合には棟単位の属性の影響を考慮できず，一方ですべて異なる棟を抽出してしまうと同じ棟での戸単位の影響を考慮できない．したがって，複数の住戸を含む棟をバランス良く選択してレコードを抽出する必要がある．そのため，まず棟単位の属性である数値コード変換住所，建物階数，築年月の和の組合せを並べ，一定間隔でその組合せに属するレコードのグループを抽出した．さらに，棟 ID の多寡はレコード同定の結果に影響を与えると考えられるため，棟 ID ありとなしのデータ比率を 1:2 として抽出した．

図 3 は物件間相違度の閾値による，再現率と適合率の関係を図化したものである．これは，ある物件間相違度の閾値以下で 2 物件が等価であると定義し，その閾値を変化させている．再現率をみると，いずれの対象地も閾値 3 で 95%を超え，閾値 3 以上での伸び率は大きくない．一方で，適合率に着目すると，いずれの対象地も閾値 6 までで 94%以上の値となり，良好な精度であるといえる．さらに，閾値 3 になると適合率は 98%を超えるが，閾値 1，2 での伸び率は大きいとはいえない．以上をふまえると，本研究で定義した物件間相違度は，閾値 3 で設定した場合に再現率，適合率ともにバランス良くレコード同定可能であることが分かった．

最後に，階層的クラスタリングの結果と実際の物件数を比較した．物件間相違度の閾値に関して，上記の 3 対象地だけで一般化することは難しいが，3 対象地が地理的，環

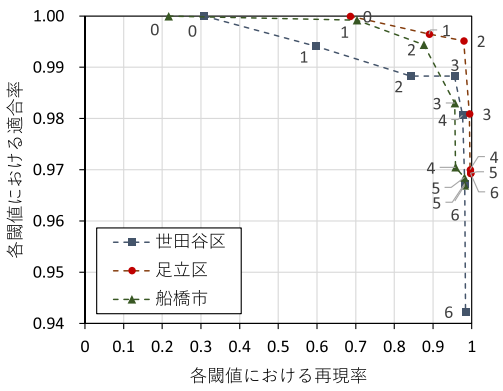

図 3　物件間相違度の閾値変化による再現率と適合率の推移

Fig. 3　Changes in the relationship between recall and precision per threshold level.

表 4　提案手法で不正解として処理された例
Table 4　Examples of prediction errors in test data.

番号	棟 ID	棟 ID 欠損	住所	物件名	築年月	建物階数	専有面積	間取り	部屋階数	部屋階数欠損	部屋番号	部屋番号欠損	件数
	同物件において提案手法で不一致（偽陰性）												
1	-	×	○	○	○	○	○	×	×	○	○	○	5
2	-	×	○	×	×	○	○	○	○	○	○	○	4
3	-	△	○	○	×	○	○	○	○	○	○	○	3
4	○	○	○	○	×	○	○	×	×	○	○	○	3
5	-	×	○	×	×	○	○	×	○	○	○	○	2
6	-	×	○	×	○	○	○	×	○	○	○	○	2
7	-	×	○	○	○	○	○	×	○	○	○	○	2
8	-	×	○	○	○	○	○	○	○	○	×	○	2
9	-	△	○	×	×	○	○	○	○	○	○	○	2
10	-	×	○	×	×	○	○	×	×	○	○	○	1
11	-	×	○	×	×	○	○	○	×	○	○	○	1
12	-	×	○	×	○	○	○	○	×	○	○	○	1
13	-	×	○	×	○	○	○	○	○	○	○	○	1
14	-	×	○	○	×	○	○	○	○	○	○	○	1
15	-	×	○	○	○	○	×	○	×	○	×	○	1
16	-	×	○	○	○	○	×	○	×	○	○	○	1
17	-	×	○	○	○	○	○	×	○	○	×	○	1
18	-	△	○	×	×	○	○	○	×	○	○	○	1
19	-	△	○	×	○	○	○	×	○	○	○	○	1
20	-	△	○	×	○	○	○	○	○	○	×	○	1
21	-	△	○	×	○	○	○	○	○	○	○	○	1
22	-	△	○	○	×	×	×	○	×	○	○	○	1
23	-	△	○	○	×	○	○	○	×	○	○	○	1
24	-	△	○	○	○	×	×	○	○	○	○	○	1
25	-	△	○	○	○	○	○	×	○	○	○	○	1
26	-	△	○	○	○	○	○	○	×	○	○	○	1
27	○	○	○	×	○	○	○	○	×	○	○	○	1
28	○	○	○	○	×	○	○	○	×	○	○	○	1
	異なる物件において提案手法で一致（偽陽性）												
1	-	×	○	○	○	○	○	○	○	○	○	○	10
2	-	×	○	×	○	○	○	○	○	○	○	○	8
3	-	×	○	○	○	○	○	○	-	△	-	△	3
4	○	○	○	○	○	○	○	○	○	○	○	○	1
5	-	×	○	△	○	○	○	○	○	○	-	△	1
6	-	×	○	○	△	○	○	○	○	○	-	△	1

表 3　正解物件数と名寄せ処理後クラスタ数の比較
Table 3　Comparison between the number of actual dwellings and that of clusters created by the proposed method.

対象地	正解物件数	名寄せ処理後クラスタ数
世田谷区	264	289
足立区	376	373
船橋市	391	392

境的にも分散しており，一定程度の信頼性が担保できると考え，クラスタリングの閾値は 3 に設定した．

表 3 をみると，正解物件数とレコード同定処理後クラスタ数の差は，世田谷区，足立区，船橋市でそれぞれ+25 件（+9.5%），−3 件（−0.8%），+1 件（+0.3%）であり，大きな誤差はみられなかった．なお，これは正解物件数とユニークであると予測された物件数の差異であり，必ずしもレコード同定の精度測定とはいえないことに留意されたい．

以上の精度検証により，提案手法の高い再現率，適合率を確認できたが，それでも一定程度のエラーは生じている．提案手法では物件名の前処理や属性相違度の導入を通して，表記揺れや集約処理の回避のためのエラーなどを一定程度許容できていると考えるが，提案手法で制御できていないエラーは，何が考えられるであろうか．正解データのうち，同物件において提案手法で不一致として処理された例（偽陰性）および異なる物件において提案手法で一致として処理された例（偽陽性）の一覧を示したものが**表 4** である．なお，欠損の有無を示すカラムは，○：両データ

に欠損なし，△：一方のデータに欠損あり，×：両データに欠損ありを示す．

同物件において提案手法で不一致の例をみると，棟 ID が欠損している場合が多く，28 例のうち 25 例を占める．さらに，物件名，部屋階数，築年月が不一致の場合についてもそれぞれ提案手法で不一致となっており，それぞれ 13 例，13 例，12 例を占める．以上のことから，同物件について提案手法で不一致となる場合は，棟 ID が欠損しているためにペナルティが加算され，加えて物件名や築年月などの属性が不一致であるために生じると考えられる．特に，物件名は表記揺れを完全にコントロールできている訳ではないため，たとえば正解：「○○アパート」，予測：「○○アパート（△△△）」や，正解：「アパート abc」，予測：「アパートエービーシー」などが不一致として処理されている例である．このような不一致は，ある程度表記揺れなどによる誤差を制御できている一方で，集約処理の回避のためのエラーを完全には制御できていないと示唆される．たとえば，不動産仲介業者が「○○アパート」を宣伝する際，物件名に部屋番号（△△△）まで付加することにより，他の物件に寄せられず追加的な情報も加えられるため，誘因となる．全体の名寄せ精度から勘案するに大きな誤差とはいい切れないものの，物件名のさらなる前処理が必要になると考えられる．また，そもそも物件名が登録されていない，すなわち欠損値である場合もあり，そのような場合は物件名を手がかりとすることはできない．

異なる物件において提案手法で一致の例では，棟 ID のみ欠損している場合が 10 件で最も多く，加えて棟 ID 欠損と物件名が異なる場合で 8 件であった．棟 ID のみ欠損の例では，部屋番号が「203」と「2030」というように異なっているものの，コサイン相違度の値が 0.3 以下であったために名寄せされてしまったと考えられる．これは，棟 ID も含めてすべての属性が一致している場合にも同様にいえる．一方，棟 ID 欠損に加えて物件名も異なる場合には，不動産仲介業者が物件名を誤って入力した可能性がある．その他，部屋階数または部屋番号が欠損している例が 4 件観測されたが，これらも情報入力者のエラーにより，適切に情報が入力されなかったと考えられる．上記の結果から，手法改善として部屋番号における属性相違度の閾値設定変更が考えられるが，部屋番号のエラーの多寡が地域によって異なると考えらえるため，今後地域を拡大した大規模データで最適な閾値を検証する必要があると考える．加えて，不動産仲介業者側の入力ミスやエラーに対しては，未入力や外れ値にエラーを返すような入力フォーマットの提供などが考えられる．

では，表記揺れが大きく，仲介業者に誘因を与える源泉となる物件名を削除して同様の名寄せ処理を行った場合，その再現率，適合率はどのように変化するのであろうか．**表 5** は提案手法において全属性を投入したものと物件名を削除したものを比較したものである．

表 5　世田谷区における提案モデルと物件名を除いたモデルの比較
Table 5　Comparison between the proposed model and a model excluding building names.

物件間相違度	全属性投入		物件名削除	
閾値	再現率	適合率	再現率	適合率
0	0.308	1.000	0.356	1.000
1	0.597	0.994	0.919	0.992
2	0.842	0.988	0.961	0.984
3	0.956	0.988	0.977	0.968
4	0.976	0.981	0.983	0.944
5	0.982	0.967	0.984	0.862
6	0.984	0.942	0.985	0.709

表 5 を観察すると，閾値の設定により精度が異なるものの，物件名を除いても再現率，適合率は大きく変化する訳ではない．また，物件間相違度の閾値が 2 までときには物件名を削除した場合の方が再現率は良好な数値となっている．しかし，閾値が 5 のときには物件名を除いた場合の適合率は 86.2%まで落ち込み，閾値が 6 のときに 70.9%となる．上記の結果から，物件間相違度の値が小さいときには物件名がノイズとなり，再現率を押し下げているものの，物件間相違度の値が大きくなると物件名が予測データの正解を高めることに貢献しているといえる．これは，集約処理の回避のためのエラーをはじめ，制御困難なエラーが，物件間相違度に影響を与えていると考えられ，不動産レコード同定特有の現象を表しているといえる．ただし，物件名を除いた場合でも閾値 3 のときに再現率 97.7%，適合率 96.8%であり，ニューラルネットワークを利用した既往研究 [16] の再現率 95%に比肩しており，十分に実用的であると考えられる．

最後に，重み付けを行った棟 ID，部屋階数，部屋番号は，提案手法の精度にどのような影響を与えているのかを検証した．特に，棟 ID は一般的な不動産データベースで必ずしも利用できる訳ではなく，その重みを 100 と 0 とで比較した．一方，部屋階数，部屋番号は一般的に入力の必要なデータであることが多く，それらの重みを 100 と 1 の場合で比較した．**表 6** は 3 属性について重みを変えて，計 4 種類のモデルの再現率，適合率をみたものである．

まず，基本モデルと棟 ID なしモデルと比較した結果，再現率に差異はないことが分かった．これにより，棟 ID の存在が過度な物件間相違度の上昇を招き，偽陰性の件数を上げている訳ではないことが明らかになった．一方で，適合率についても棟 ID の有無によってその値が大きく変わるわけではないが，閾値が高くなっても基本モデルでは適合率が維持される傾向にあると分かった．これは，棟 ID が間違って同じ棟としてリンクされることを防ぎ，結果的に偽陽性の件数が抑えられていると推察される．

続いて，部屋階数と部屋番号の重みがそれぞれ 1 である

表 6　世田谷区における提案モデルと棟 ID，部屋階数，部屋番号の重みを変化させたモデルの比較

Table 6　Comparison between the proposed model and models changing the weights of building ID, number of floor, and room number.

棟 ID：部屋階数：部屋番号の重み付け比率	100:100:100（基本モデル）		0:100:100		100:1:1		0:1:1	
閾値	再現率	適合率	再現率	適合率	再現率	適合率	再現率	適合率
0	0.308	1.000	0.308	1.000	0.308	1.000	0.308	1.000
1	0.597	0.994	0.597	0.994	0.597	0.589	0.597	0.589
2	0.842	0.988	0.842	0.988	0.845	0.453	0.845	0.453
3	0.956	0.988	0.956	0.988	0.961	0.285	0.961	0.285
4	0.976	0.981	0.976	0.981	0.984	0.181	0.984	0.181
5	0.982	0.967	0.982	0.965	0.996	0.137	0.996	0.137
6	0.984	0.942	0.984	0.920	0.998	0.109	0.998	0.109

モデルを比較した結果，棟 ID の有無にかかわらず，適合率が低くなる傾向であると分かった．このように棟 ID で大きな変化がなく，部屋階数，部屋番号で変化が見られたのは，データの階層性が関連していると考えられる．今回用いた検証用のレコードは 999 件あり，そのうち棟 ID は 75 棟，部屋階数は 150 階しか識別できないが，部屋番号は 264 戸をユニークに識別できる．そのため，部屋番号は，異なる物件を誤ってリンクさせてしまうことを防ぎ，適合率の寄与に大きく関与すると考えられる[*18]．

このように，棟 ID の影響や部屋番号の重要さについて明らかになってきたものの，本検証では前処理を十分に行っており，かつ世田谷区を検証対象地として質の高いデータを用いている．そのため，前処理が十分でないケースや，部屋番号などの属性に欠損が多いと考えられる地方部でのケースでも，今後検証を進める必要がある．

6. おわりに

本研究では，不動産レコード同定の枠組みにおいて，物件名，建物属性，棟レベルの識別 ID などについてデータの前処理を行い，物件間相違度を各属性相違度を用いて測定し，階層的クラスタリングによって不動産情報の集約を行う一連の手法を提案した．加えて，集約された物件情報が実用的な精度を担保できるかについて検証を行った．

世田谷区，足立区，船橋市で精度検証を行った結果，いずれの対象地でも再現率 95%，適合率 94%を超え，不動産情報を扱うにあたって実用的な精度を達成することができた．本研究では主にルールベースで手法を構築し，物件名を除いた場合であってもニューラルネットワークを利用した既往研究と同程度の再現率を得ており，本研究は簡便な方法で高い精度に至ったことを示している．

さらに，同物件において提案手法で不一致であったレコードは，棟 ID の欠損に加えて一部属性が不一致であるために生じる場合が多いと分かった．再現率や適合率の高さも加味すると，作業ミスや表記揺れ，集約処理の回避によるエラーなどを許容して集約ができているといえるが，それでも一部で上記エラーを制御できない例が存在すると示唆された．一方，異なる物件において提案手法で一致していたレコードは，部屋階数や部屋番号の欠損，部屋番号のわずかな表記揺れによって引き起こされることが分かった．

上記をふまえ，物件名や棟 ID を削除したり，部屋階数，部屋番号の重みを変化させたりしてレコード同定を行ったところ，棟 ID が適合率の安定性に寄与していること，部屋番号が手法の精度に大きな影響を与えることが明らかになった．これらは，それぞれの階層（棟 ID であれば棟，部屋番号であれば戸）で異なる物件間で誤ってリンクすることを防ぐ役割を持つと考えられる．特に，部屋番号は唯一住戸をユニークに識別できるため，モデル間の適合率に大きな影響を与えたといえる．

今回提案した集約手法では比較的良好な精度を達成したものの，以下のような限界が存在する．まず，今回検証を行ったのは東京都区部および近郊市にあたる地域であり，比較的情報が充実していると考えられる．そのため，都市圏外などで同種の手法を展開する際には，欠損値の存在から精度が低下すると考えられる．このように，本研究の検証地域は限定的であり，教師データがパラメータを大域的に推定できるほど十分でなく，教師なしで名寄せモデルを構築する必要があった．そのため，提案手法は実務的な知見に基づき，現実的に妥当と思われる属性相違度の閾値を設定した．今後，最適なパラメータ設定は大規模データからグリッドサーチを用いた交差検証などで行う必要がある．また，網羅性の観点からみると，不動産募集広告が存在しない住宅は必然的にデータベースに含まれないことになる．住宅ストックの把握などへのデータベース利用に対しては，住宅市場に出現しない物件を補完するための技術についても検討が必要である．さらに，本研究では戸寄せに

[*18] なお，部屋階数のみ重みを 100 としたモデルでは，クラスタリングの閾値 3 で 43.4%，閾値 4 で 36.2%と相対的に低い適合率となっていることからも，部屋階数でなく部屋番号が重要な属性であるといえる．

特化しており，棟寄せを同時に行うような手法とはなっていない．棟寄せを行う際には棟だけよりも戸の情報を合わせることで棟寄せ精度を改善できる可能性があるため，先に棟単位で寄せてから戸の情報をマッチングさせ，棟寄せの精度をあげるような2段階での手法を想定し，今後の改善を課題としたい．以上の限界を考慮しても，本研究で提案した手法は，その特性を理解したうえで地域や用途を限定することで，十分に利用価値のあるものであるといえる．

レコード同定の対象としての不動産は，2章にて指摘したとおり，「特定性」「階層性」「市場性」「情報非対称性」など，他の対象にはない独特の特徴を有することから，今後さらに研究を深掘りしていく価値があると考えている．特に，近年の不動産募集広告において充実が著しいマルチメディア情報，特に間取り図や室内・外観写真，パノラマ写真，動画などは，レコード同定において有力な手がかりとなりうる．マルチメディア情報は，不動産が「特定性」という特徴を有するからこそ利用可能なものである．

今後，これらのマルチメディア情報を特徴量として用いて精度向上を図るとともに，小地域レベルでの価格指数の構築など，応用にも積極的に還元していく予定である．

謝辞　本研究はJSPS科研費20H00082および20K14898の助成を受けたものです．

参考文献

[1] Shimizu, C., Nishimura, K.G. and Asami, Y.: Search and Vacancy Costs in the Tokyo housing market: Attempt to measure social costs of imperfect information, *Review of Urban & Regional Development Studies*, Vol.16, No.3, pp.210–230 (2004).

[2] Shimizu, C., Nishimura, K.G. and Watanabe, T.: House prices at different stages of the buying/selling process, *Regional Science and Urban Economics*, Vol.59, pp.37–53 (2016).

[3] Ohnishi, T., Mizuno, T., Shimizu, C. and Watanabe, T.: Power laws in real estate prices during bubble periods, *International Journal of Modern Physics: Conference Series*, Vol.16, pp.61–81, World Scientific (2012).

[4] 不動産情報サイト事業者連絡協議会：不動産XML研究 (2002)，入手先 ⟨https://www.rsc-web.jp/kat/xml/index.html⟩.

[5] Dunn, H.L.: Record Linkage, *American Journal of Public Health and the Nations Health*, Vol.36, No.12, pp.1412–1416 (online), DOI: 10.2105/AJPH.36.12.1412 (1946).

[6] Diewert, W.E., Nishimura, K.G., Shimizu, C. and Watanabe, T.: *Property Price Index*, Advances in Japanese Business and Economics, Springer (2020).

[7] Block, R.J.: Authority Control: What It Is and Why It Matters (2016), available from ⟨https://studylib.net/doc/9764486/authority-control–what-it-is-and-why-it-matters⟩.

[8] Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P.: Automatic linkage of vital records, *Science*, Vol.130, No.3381, pp.954–959 (online), DOI: 10.1126/science.130.3381.954 (1959).

[9] Fellegi, I.P. and Sunter, A.B.: A Theory for Record Linkage, *Journal of the American Statistical Association*, Vol.64, No.328, pp.1183–1210 (online), DOI: 10.1080/01621459.1969.10501049 (1969).

[10] Wilson, D.R.: Beyond Probabilistic Record Linkage: Using Neural Networks and Complex Features to Improve Genealogical Record Linkage, *International Joint Conference on Neural Networks*, pp.9–14 (2011).

[11] Cucerzan, S.: Large-Scale Named Entity Disambiguation Based on Wikipedia Data, Technical Report (2007).

[12] Ono, S., Sato, I., Yoshida, M. and Nakagawa, H.: Person name disambiguation in Web pages using social network, compound words and latent topics, *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), Vol.5012 LNAI, Springer, Berlin, Heidelberg, pp.260–271 (online), DOI: 10.1007/978-3-540-68125-0_24 (2008).

[13] Dumont, B., Maggio, S., Said, G.S. and Au, Q.-T.: Who wrote this book? A challenge for e-commerce, *Proc. 5th Workshop on Noisy User-generated Text* (*W-NUT 2019*), pp.121–125, Association for Computational Linguistics (online), DOI: 10.18653/v1/D19-5516 (2019).

[14] Harrison, D. and Rubinfeld, D.L.: Hedonic housing prices and the demand for clean air, *Journal of Environmental Economics and Management*, Vol.5, No.1, pp.81–102 (online), DOI: 10.1016/0095-0696(78)90006-2 (1978).

[15] Lee, S.: Distance and diversification, *Journal of European Real Estate Research*, Vol.9, No.2, pp.183–192 (online), DOI: 10.1108/JERER-02-2016-0010 (2016).

[16] 門　洋一，広方　崇，松村浩二，汪雪テイ，山崎俊彦：ニューラルネットワークを利用した集合住宅の物件情報の名寄せ，第34回人工知能学会全国大会（JSAI 2020）予稿集，1N5-GS-13-03（オンライン），DOI: 10.11517/pjsai.JSAI2020.0_1N5GS1303 (2020).

馬場 弘樹

1987年生．2011年東京大学工学部都市工学科卒業．2013年同大学大学院工学系研究科都市工学専攻修士課程修了．2019年同博士課程修了．2019年から東京大学空間情報科学研究センター特任研究員を経て2020年より特任助教．不動産データベースの構築や空き家の空間分布推定等の研究に従事．2019年より麗澤大学AI・ビジネス研究センター客員准教授，2020年より同都市・不動産研究センター客員准教授を兼務．

関口 知子

1988 年埼玉大学理学部数学科卒業．銀行の情報システム開発業務，人工知能開発サービス提供企業における自然言語処理・音声解析処理の研究開発業務への従事等を経て，2010 年（株）リッテルに入社．2011 年より（株）LIFULL で，各種データ解析業務に従事．2020 年より東京大学空間情報科学研究センター協力研究員を兼務．

清田 陽司　（正会員）

1975 年生．1998 年京都大学工学部電気工学第二学科卒業．2000 年同大学大学院情報学研究科修士課程修了．2004 年同博士課程修了．2004～2012 年まで東京大学情報基盤センター助手・助教・特任講師．2007 年に東京大学発スタートアップ（株）リッテルを共同創業し，企業買収を経て 2011 年より（株）LIFULL 主席研究員．不動産分野における AI 技術全般の研究開発，および共同研究やデータセット提供等の産学連携に従事．本学会 DBS 研究会運営委員，UBI 研究会幹事，人工知能学会編集委員長等を担当．人工知能学会，言語処理学会，日本データベース学会，日本不動産学会各会員．東京大学空間情報科学研究センター客員研究員等を兼務．

清水 千弘

日本大学スポーツ科学部教授，東京大学空間情報科学研究センター特任教授および麗澤大学 AI ビジネス研究センターセンター長．マサチューセッツ工科大学研究員を兼務する．東京大学博士（環境学）．麗澤大学経済学部，シンガポール国立大学不動産研究センター教授を経て現職．専門はビッグデータ解析・経済測定．

（担当編集委員　松村 敦）