キャッチコピーを用いた
飲食店用不動産賃料推定モデルの改善

Improvement of Rent Estimation of Real Estate for Restaurants Using Catchphrase

鶴山優季子 [1], 諏訪博彦 [1], 小川祐樹 [2], 荒川豊 [1], and 安本慶一 [1]

[1] 奈良先端科学技術大学院大学　Nara Institute of Science and Technology
[2] 立命館大学　Ritsumeikan University

概 要

飲食店向け不動産物件の賃料は，ベテラン営業職員が培ってきた経験や勘といった暗黙知に基づいて決定されている．賃料の決定要因として，物件固有の情報である静的情報，物件周辺の情報である動的情報，物件の特徴や雰囲気などを含む潜在的情報の 3 つが挙げられている．先行研究では，指標化の難しい潜在的情報にキャッチコピーを用いている．しかし，キャッチコピーの文脈が考慮されていないなどの課題がある．そこで本研究では，Doc2Vec によりベクトル化したキャッチコピーを用い，重回帰分析とランダムフォレストによりモデルを構築した．その結果，重回帰分析を用いた場合に決定係数が高くなり，また静的情報，動的情報，潜在的情報の 3 つの要因を全て用いた場合に，決定係数が 0.650 と最も高くなるという結果が得られた．

キーワード：機械学習，データマイニング，賃料推定，センシング，自然言語処理

Abstract

The rent of real estate for restaurants is determined based on tacit knowledge such as experiences or intuitions gained by veteran sales man. Determinants of rent include static information, specific to the properties, dynamic information, around the properties, and latent information, including characteristics or atmosphere. We used catchphrases vectorized by Doc2Vec for latent information which is difficult to index, and constructed the models by linear regression and random forest. As a result, using linear regression, the coefficient of determination became high, and using all three factors of static, dynamic, and latent information, the coefficient of determination became the highest.

Keywords: Machine Learning, Data Mining, Rent Estimation, Sensing, Natural Language Processing

1　はじめに

機械学習の発展に伴い，不動産分野においても住宅物件の価格推定 [1][2][3] などの営業支援が活発に行われている．一方で，飲食店を対象とした不動産に限ってみると，経験豊富なベテラン営業職員が培ってきた経験や勘といった暗黙知に基づいて賃料を決定している現状がある．この手法では，賃料を決定している要因が明確ではなく，人により賃料が異なるといった問題や，新人営業職員への知識継承が難しいといった問題が生じる．飲食店不動産の選定には，ガス，排気設備，物件周辺の通行量，視認性などの固有の属性が存在する．そのため，立地や間取りが選定の中心である従来の住宅系の価格推定手法をそのまま用いることはできない．こういったことから，飲食店向け不動産の賃料は，ベテラン営業職員の暗黙知により決定されている．

荒川らの先行研究 [4] では，飲食店向け不動産物件の賃料に影響を与えている暗黙知に基づいた賃料推定モデルを提案している．この賃料推定モデルは野中らの SECI モデル [5] をベースに構築されており，ベテラン営業職員に対して行った詳細なインタビューから暗黙知を表出化している．これらの暗黙知は，物件固有の情報である静的情報，物件周辺の日々変化する動的情報，物件の雰囲気や特徴を含む潜在的情報の 3 つに分けられ，指標化されている．潜在的情報は，静的情報や動的情報と比べ指標化の難しい情報を含んでいる．先行研究では，潜在的情報として各物件に与えられているキャッチコピーを用いている．しかしその手法は，キャッチコピーに含まれている単語の正負判定に止まっている．この手法では，キャッチコピーの文脈を考慮することはできず，また同一

の単語でなければ判定することができないといった問題が生じる．そのため，キャッチコピーの文脈や曖昧さを考慮することが課題となっている．

本研究では，飲食店向け不動産物件の賃料推定システムを構築することを目的とする．モデルの構築は，荒川らの既存モデル [4] に基づいて行う．また課題を解決するため本研究では Doc2Vec を用いる．Doc2Vec は任意の長さの文書をベクトル化する技術で，文書内の単語の意味的表現も学習することができる [6]．本研究では，Doc2Vec によりベクトル化されたキャッチコピーを潜在的情報として用いる手法を提案し，その評価を行う．モデルの構築には，重回帰分析とランダムフォレストを用い，交差検証により比較を行う．データセットは，契約が成立した 193 件の飲食店向け物件を対象とした．その結果，重回帰分析を用いた際に決定係数が 0.650 となり，検証した中で最も高い値が得られた．

本稿の構成について述べる．2 章では賃料推定に関する研究について述べる．3 章では本研究で提案する賃料推定モデルについて述べる．4 章では 3 章で構築したモデルの評価を行い，5 章では考察，6 章では結論を述べる．

2　関連研究

不動産価格の推定に関する研究を述べる．三浦ら [1] は，不動産側のデータベースと，インターネット上にある膨大な情報から対象エリアの評判に関する変数を抽出し，それらを組み合わせる方法を提案している．評判に関する変数はインターネット上で地名を検索し，形態素解析を行うことで抽出されている．モデルの構築にはヘドニック分析法が用いられ，不動産側のデータベースのみの場合は決定係数が 0.528，インターネット上の情報を組み合わた場合は決定係数が 0.560 という結果が得られている．

Wu ら [2] は台湾での住宅選定に影響があるという風水に着目している．モデルの構築にはバックプロパゲーションニューラルネットワーク，ファジーニューラルネットワーク，独自に開発したハイブリッド遺伝ベースのサポートベクター回帰からなる複数のアルゴリズムを用いて比較を行っている．その結果，いずれの手法においても，風水を考慮した方がより推定精度が良いという結果が得られている．

Chiarazzo ら [3] はイタリアにおいて，交通システムと地域ごとの環境の質が不動産価格に影響を与えていると考え，人口ニューラルネットワークを用いて検証している．その結果，全 42 ある属性の内の 8 番目に環境汚染に関する属性が挙げられ，また交通に関する属性も 15 番目付近に位置していることを明らかにしている．

このように，賃料推定に関する研究は多数存在する．しかしこれらの関連研究では，対象物件が一般住宅であり，選定の基準が異なる飲食店向けの不動産には同じ手法を適用することはできない．本研究では，これらの研究を参考にしながら，飲食店向けの賃料推定モデルを構築する．

3　既存モデルと改善モデル

本研究では，荒川ら [4] の先行研究に基づき，飲食店向けの賃料推定モデルを構築する．本章では既存モデルの概要と課題，課題を解決するための手法について述べる．

3.1　既存モデルと課題

本節では，既存モデルの概要と用いた特徴量，また構築されたモデルの課題について述べる．

3.1.1　既存モデル概要

先行研究 [4] では，野中ら [5] の SECI モデルに基づき，賃料推定モデルを構築している．SECI モデルは，暗黙知を伝承するための知識創造モデルである．暗黙知とは言語化できない知識を，形式知とは言語化できる知識のことを意味する．SECI モデルと先行研究の関係を図 1 に示す．野中らによると，知識創造は（1）共同化→（2）表出化→（3）連結化→（4）内面化→（1）共同化といったサイクルを繰り返すことで可能となる．SECI モデルは，知識は暗黙知を表出化して形式知にし，連結化することで概念

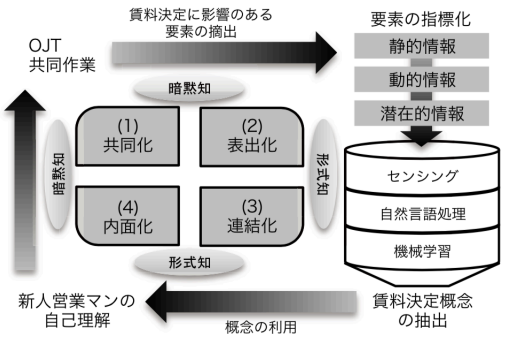

図 1: SECI モデルと先行研究の関係 [4]

として共有・伝承が可能になることを示している．このモデルに基づくと，ベテラン営業職員の暗黙知は，表出化して形式知にし，連結化することで共有・伝承が可能となる．

そこで先行研究では SECI モデルに基づき，暗黙知を形式知として表出化するためベテラン営業職員に対してインタビュー調査を行った．インタビューに基づき，特徴量の取捨選択を行った結果，賃料の決定要因としては，物件固有の情報である静的情報，物件周辺の日々変化する動的情報，物件の特徴や雰囲気を含む潜在的情報の 3 つが抽出されている．

3.1.2　既存モデルの特徴量

本項では，既存モデルの特徴量について述べる．

(1) 静的情報

静的情報とは物件固有の情報であり，年数が変化しても基本的には変化しない情報のことを指す．インタビューにより，具体的な静的情報として坪数，駅からの距離，階数，居抜きが挙げられた．居抜きとは物件に付帯しているテーブル，カウンター，ガスレンジなどの設備を指しており，居抜きであれば 1，そうでなければ 0 として指標化している．これらの情報は不動産会社の Web ページ上に掲載されており，改めて指標化を行う必要はない．

(2) 動的情報

動的情報とは物件周辺の日々変化する情報のことを指す．インタビューにより，具体的な動的情報として地域ポテンシャル，通行量，視認性が挙げられた．地域ポテンシャルは，最寄駅の平均坪単価に対象物件の坪数をかけた値と定義している．また通行量と視認性はベテラン営業職員 2 名による 5 段階評価の平均値とし，さらに通行量と視認性の積を取ることで，対象物件の見つけやすさを表す指標としている．積を取る理由は，例えば視認性は高いが通行量が全くないといった場合に見つけやすさを低く評価したいからである．

(3) 潜在的情報

潜在的情報とはベテラン営業職員による指標化が難しい情報を指す．既存モデルでは，潜在的情報として，物件に付与されたキャッチコピーを用いた．具体例としては,「大通り交差点すぐそば！」,「焼肉店居抜き店舗」,「視認性良好、角地店舗です」などがある．これらのキャッチコピーに含まれる名詞，形容詞を形態素解析により求め，それらの単語が賃料に与える影響を求めている．具体的には，賃料に正の影響を与える単語として「大通り」，また負の影響を与える単語として「焼肉」,「角地」などが得られている．

3.1.3　既存モデルの課題

既存モデルでは，キャッチコピーを用い，文中に含まれる名詞や形容詞が賃料に対して正負どちらの影響を与えるかを検証している．しかしこの手法ではキャッチコピーの文脈を考慮することができず，同一の単語でなければ検出することができない．また全ての単語についての正負判定が必要といった問題が生じる．そのため，キャッチコピーの文脈や曖昧さを考慮できるモデルの構築が必要である．

3.2　改善モデルの提案

本節では，既存研究の課題を解決するため，新たなモデルの構築を提案する．

3.2.1　Doc2Vec

本研究においても，潜在的情報の情報源としてキャッチコピーを用いるが，文脈や単語の曖昧さを考慮するため，Doc2Vec を用いた手法を提案する．

Doc2Vec とは，機械学習を用いた，任意の長さの文書をベクトル化できる技術である [6]．類似手法である Word2Vec は，具体例として「王様」－「男＋「女」＝「女王様」といった概念の足し引きができる．Doc2Vec は，この Word2Vec の拡張版であり，単語の意味的表現を文書レベルで学習することができる．言語をベクトル表現で扱うことができるため，コサイン類似度を算出することができ，またその類似度から文書の分類を行うこともできる．

以上のような Doc2Vec の技術を用いることで，既存モデルで考慮しきれなかった文脈や曖昧さについてもモデルに組み込むことができると考える．

3.2.2　Doc2Vec による潜在的情報のモデルの構築

本研究では，24000 件以上の物件のキャッチコピーを用いて，Doc2Vec によるキャッチコピーのベクトル化を行った．キャッチコピーは，あらかじめ文章を単語ごとに分割する形態素解析の処理を行い，Doc2Vec の実施には gensim ライブラリを使用した．

Doc2Vec によるキャッチコピーのモデル構築を行う際のパラメータとして，ベクトルの次元数，ウィンドサイズ，最小カウント数の 3 つを使用した．ウィンドサイズは，現在の単語と予測する単語の最大の距離であり，最小カウント数は，出現回数が低いものを無視する閾値である．キャッチコピーにとって最適なパラメータを決定するため，ウィンドサイズを

1, 2, 3 の 3 種類，最小カウント数は 5, 10, 15 の 3 種類の計 9 種類の組み合わせを用い，予備解析を行った．その結果，ウィンドサイズが 3，最小カウント数が 10 の場合に決定係数が高い結果となった．このことから，以降ではこれらの値を使用し，モデルを構築することとする．またベクトルの次元数により推定結果が大きく異なることが予想されるため，次元数を 1，5，10，50，100 の 5 種類とし，比較することとする．

3.3 賃料推定モデルの構築

指標化された構成要素と賃料の関係を連結化するため，機械学習を用いた賃料推定モデルの構築を行う．本研究では，機械学習法として重回帰分析（Linear Regression，以下 LR）とランダムフォレスト（Random Forest，以下 RF）の 2 種類を用いてモデルを構築する．モデルの構築に用いる要因は，静的情報，動的情報，潜在的情報の 3 つであり，具体的には静的情報として坪数，駅徒歩時間，居抜き，階数，動的情報として地域ポテンシャル，通行量×視認性，潜在的情報としてベクトル化されたキャッチコピーを用いる．

賃料推定モデルの精度を評価するため，決定係数と平均二乗誤差を用いる．決定係数は，当てはまりの良さを表し，最も良いスコア値は 1.0 である．それに対し平均二乗誤差は，真値と推定値がどれほど乖離しているかを表しており，0 に近いほど優れていることを示す．

4 賃料推定モデルの評価

提案した賃料推定モデルの評価を行うため，機械学習法別，構成要因別，また潜在的情報の組み込み方による精度の比較を行う．

4.1 機械学習法の比較

LR と RF により検証した評価結果を表 1 に示す．ここでは，静的情報，動的情報，潜在的情報をモデルの構成要因とし，検証には 3-fold 交差検証を用いた．表 1 には，ベクトル化した潜在的情報の次元数，機械学習法別の決定係数，平均二乗誤差を示している．

決定係数について機械学習法別に比較すると，ベクトルが低次元の場合は LR，ベクトルが高次元の場合は RF が高いという結果が得られた．同様に，平均二乗誤差について機械学習別に比較すると，ベクトルが低次元の場合は LR，高次元の場合は RF が低い値を得た．これらのことから，ベクトル数が低次元の場合は LR，高次元の場合は RF が適しているという結果が得られた．しかし次元数で比較すると，高次元になるにつれ，LR，RF 共に決定係数が下がり，平均二乗誤差も大きくなることがわかった．このことから，高次元は提案モデルに適していないことが考えられる．以降では，キャッチコピーの次元数は低次元である 1，5，10 の 3 種類を用い，低次元でのモデル構築に適している LR を用いて検証を行う．

表 1: 機械学習法別の評価結果

次元数	決定係数		平均二乗誤差	
	LR	RF	LR	RF
1	0.650	0.557	93364	100962
5	0.610	0.544	89491	102359
10	0.600	0.457	94390	109209
50	0.155	0.412	126702	108282
100	-1.259	0.370	167152	121236

4.2 構成要因の比較

本研究では，潜在的情報であるベクトル化を行ったキャッチコピーが賃料に及ぼす影響について検証するため，構成要因を組み合わせたモデルを構築する．データセットは，表 2 のような組み合わせとする．モデルの構築には，LR を用い，3-fold 交差検証を用いて比較を行った．

各データセットごとの決定係数と平均二乗誤差を表 3 に示す．表 3 より，決定係数が最も高くなったのはデータ番号 11 の静的・動的・潜在的（次元数 1）であることがわかる．しかし潜在的情報を組み込んでいないデータ番号 4 と比べると決定係数に大きな差はなく，また平均二乗誤差に関してはデータ番号 4 の方が小さな値となった．また潜在的情報のみのデータセットであるデータ番号 1 から 3 より，潜在的情報のみでは賃料の推定が難しいという結果が得られた．

4.3 潜在的情報の組み込み

本節では潜在的情報を 2 種類の方法を用いてモデルに組み込み，その比較を示す．一つは静的・動的・潜在的情報の組み合わせでモデルを構築する方法，も

表 2: データセットの組み合わせ

データ番号	組み合わせ
1	潜在的 (次元数 1)
2	潜在的 (次元数 5)
3	潜在的 (次元数 10)
4	静的・動的
5	静的・潜在的 (次元数 1)
6	静的・潜在的 (次元数 5)
7	静的・潜在的 (次元数 10)
8	動的・潜在的 (次元数 1)
9	動的・潜在的 (次元数 5)
10	動的・潜在的 (次元数 10)
11	静的・動的・潜在的 (次元数 1)
12	静的・動的・潜在的 (次元数 5)
13	静的・動的・潜在的 (次元数 10)

表 3: データセット別の評価結果

データ番号	決定係数	平均二乗誤差
1	-0.031	152777
2	-0.069	143466
3	-0.192	164440
4	0.646	89671
5	0.290	127260
6	0.204	119232
7	0.161	128156
8	0.610	90990
9	0.571	89756
10	0.575	92753
11	0.650	93364
12	0.610	89491
13	0.600	94390

う一つは静的・動的情報でモデルを構築した後，真値と推定値の残差を潜在的情報により埋め合わせるという手法である．

後者の手法による推定結果を表 4 に示す．本手法では，モデルの構築に重回帰分析を用い，3-fold 交差検証により検証を行った．また，表に示した決定係数，平均二乗誤差は，静的・動的情報で推定した賃料と潜在的情報で推定した残差の和を真値と比較したものである．

まず，静的･動的情報を用いて重回帰分析によりモデルを構築した場合，決定係数は 0.646 となる（表 3）．その値と比較すると，表 4 における決定係数は，全ての次元数において低くなっている．また高次元になると，推定賃料が負の値をとる場合もあった．そ

表 4: 潜在的情報による残差の埋め合わせ

潜在的情報の次元数	決定係数	平均二乗誤差
1	0.637	81556
5	0.625	85972
10	0.619	85561

れらのことから潜在的情報による埋め合わせの効果はあまりないと考えられ，静的・動的・潜在的情報の 3 つの要因を組み合わせたモデルが有効であると言える．

5 考察

表 3 のデータ番号 1 から 3 を見ると，ベクトル化を行ったキャッチコピーのみによる賃料の推定は，難しいということがわかる．データセットの中には，異なる物件であっても全く同じキャッチコピーを使用している物件もあり，潜在的情報のみでは，物件固有の情報などを汲み取ることが難しかったと考えられる．また潜在的情報は，他の情報と組み合わせることで，決定係数が高くなるという結果が得られた．これはキャッチコピーで汲み取ることが難しかった情報を組み込んだため，精度が向上したと考えられる．

また，決定係数に着目すると，いずれの組み合わせにおいても次元数が 1 の場合に最大のスコア値を取っている．また平均二乗誤差に着目すると，いずれの組み合わせにおいても次元数が 5 の場合に最小のスコアを取っている．このことから，キャッチコピーの次元数は 1 から 5 の間とするのが最適であると考える．今回キャッチコピーのモデルを構築した際に使用したデータの中には，短いもので 3 単語からなるキャッチコピーを確認しており，4 単語からなるキャッチコピーも複数確認している．このことから，キャッチコピーのベクトルは低次元の場合が適していると考える．

6 おわりに

本研究では，飲食店向け賃料推定システムの構築に向け，機械学習法を用いてモデルの検証を行った．機械学習法には，重回帰分析とランダムフォレストを用い，構成要因として静的情報，動的情報，潜在的情報を用いた．その結果，重回帰分析を用いた場合に決定係数が 0.650 と高くなり，平均二乗誤差が 93364 と低くなった．

また，構成要因別に比較を行ったところ，静的，動的，潜在的情報の 3 つの要因を全て組み合わせることで，決定係数が最も良い結果となった．さらに，潜在的情報は，1, 5, 10, 50, 100 次元にベクトル化を行ったが，高次元の場合は LR，RF 共に決定係数が低くなったため，低次元によるモデル構築が適しているという結果を得た．

今後の課題としては，推定精度の向上のため，新たな変数の組み込み，新たなモデルの構築などが挙げられる．

参考文献

[1] Takahumi Miura, and Asami Yasuhi: Hedonic Analysis for Estimation of Condominium Rent Utilizing WEB Information, Prosedia-Social and Behavioral Science 21, pp.147-156, 2011.

[2] Wu, C. H., Li, C. H., Fang, I. C., Hsu, C. C., Lin, W. T., and Wu, C. H.: Hybrid genetic-based support vector regression with shui theory for appraising real price, First Asian Conference on Intelligent Information and Database Systems, pp.295-300, 2009.

[3] Chiarazzo, V., Caggiani, L., Marinelli, M., and Ottimanelli, M.: A Neural Network based model for real estate price estimation considering environmental quality of property location, Transportation Research Procedia, Vol.3, pp.810-817, 2014.

[4] 荒川周造，諏訪博彦，小川祐樹，荒川豊，安本慶一，太田敏澄：暗黙知に基づく飲食店向け不動産賃料推定モデルの提案，情報処理学会論文誌，Vol.59, No.1, pp.33-42, 2018.

[5] Nonaka, I., and Takeuchi, H.: The knowledge creation company: how Japanese companies create the dynamics of innovation, 1995.

[6] https://deepage.net/machine_learning2017/01/08/doc2vec.html，最終閲覧日 2019-1-23.