【原著論文】

位置と名称情報を持つ店舗・事業所データの時空間結合手法の開発 －都市地域分析への応用に向けて－

秋山祐樹*・柴崎亮介**

Spatio-temporal Integration Method for Shop and Office Data with Location Information and Application for Urban and Regional Analysis

Yuki Akiyama, Ryosuke Shibasaki

Abstract: A method of spatio-temporally integrating shop/office information including names and locations is developed and applied to whole Japan. A typical example of such shop/office information is a telephone directory. If we can integrate these data of different years spatio-temporally, i.e. identify whether a shop/office at a specific location in a specific year is identical or not with a shop/office in the different years at the same location, we can map the changes and continuation of shops/offices in an entire city or region. The spatio-temporal integration method includes; 1) a geo-coding method at the floor level of a building, 2) a method of identifying shop/office names considering geographic names that appears very frequently at a specific local area. The method is validated in several study areas and finally applied to entire Japan.

Keywords: 時空間結合 (spatio-temporal integration), 自然言語処理 (natural language processing), 単語の同一性評価 (identify verification of word), 空間インデックス (spatial index), 頻出語 (frequently-appearing word)

1. はじめに

事業所・店舗等の立地分布の変化は都市や地域の活力や魅力に大きな影響を与える。その動向を可能な限り詳細かつ定量的に明らかにすることは、都市・地域空間の持続的な再生に不可欠である。しかし多くの一般的な地域情報は市区町村・町丁目単位等で集計された統計情報であり、個店あるいは商店街レベルの詳細な変化を把握することは困難である。

一方、個々の店舗や事業所の詳細な位置や名称は電話帳や商店会の会員リスト等から得ることが出来る。複数年次のこうしたデータを突き合わせ、ある場所、ある年次に存在した店舗や事業所が、他の年次に同じ場所に存在しているのか、他の店舗等へと変化しているのかなどを知ることが出来る。

このような店舗・事業所単位の詳細な時系列変化データがあれば、例えば近年注目されている「フードデザート問題」や「まちづくり」に取り組む際に、有用な基礎的データとなる。また店舗・事業所の時

系列変化情報として事業所・企業統計調査が利用出来るが、本統計は5年に1度しか更新が行われない上に、集計単位は最小で町丁目等单位である。店舗・事業所単位の時系列変化データはこうした既存統計の欠点の補完的役割を果たす可能性も秘めている。

1. 1. 既存研究

位置や名称付きの店舗・事業所等の情報の異なる年次同士の突き合わせは、これまでも都市・地域の変容に関する調査・研究の中で頻繁に行われてきた。

例えば阿藤ほか(2006)は首都圏郊外の鉄道駅周辺商業集積を対象に、位置や名称付きの店舗・事業所情報から新規出店と撤退廃業店舗の立地特性を分析し、駅前地区の変容実態と課題を明らかにしている。アンケート調査によりデータ収集を行った例としては浅野(2002)があり、北陸甲信越地方の地方都市中心部にある大規模店舗の閉店や移転の情報を、各自自治体の商業担当課へのアンケートにて収集した。こ

* 正会員 東京大学空間情報科学研究センター (The University of Tokyo)
〒277-8568 千葉県柏市柏の葉5-1-5
E-mail: aki@iis.u-tokyo.ac.jp

** 正会員 東京大学空間情報科学研究センター (The University of Tokyo)

の研究では位置や名称付きの店舗・事業所等の情報の時空間結合は手作業で行われ、确实・詳細な変化検出が出来る反面、多量のデータの処理には適さず、例えば日本全国に毎年適用することは難しい。

多量のデータの時空間結合を情報技術により自動化するためには住所等を緯度経度等に変換し地図に落とす技術、特定の場所に存在する店舗や事業所を名称突き合わせにより同一性判定をする技術を開発する必要がある。住所を緯度経度に変換する技術はアドレスマッチングとして知られており、すでに商用化・公開されている。しかし現在我が国で最も詳細なアドレスマッチングも住居表示における号レベルであり、建物名称や階数・部屋番号が住所に記載されていても、いわゆるランドマーク（例えば六本木ヒルズ等の著名建物）を除いて建物・階レベルまで位置を落とし込む機能は実現していない。店舗等の変化を捉えたり、空室状況を推定したりするためには階数・部屋レベルの位置同定が有利である。

一方、伊藤・曲淵（2001）は異なる時点間の電子住宅地図、すなわち建物ごとに店舗・事業所名称を突き合わせる時空間結合技術を開発した。しかしこの手法は特定の種類の電子地図への適用に留まっており、異種データの結合には適用されていない。また時空間統合処理は東京23区内が対象であり、後述する雑音語処理やローカル頻出語に見られる課題への対応は困難であると見られ、そのままでは広域への適用には限界があるものと考えられる。

1.2. 目的

本研究では施設や店舗等の名称情報と住所情報からなるデータを対象に、年次が異なるデータを空間的に重ね合わせ、さらに名称を突き合わせて店舗等の同一性判定を行い、店舗・事業所一件一件の時系列変化情報を明らかにする手法を日本全国に適用することを念頭に開発し、それを実装した自動処理システムを実現する。本研究ではこの処理のことを「時空間結合」と呼ぶ。名称情報と住所や経緯度がセットになったデータは数多い。最も大規模で網羅性が高いデータの例としては住宅地図の入居者データ、電話帳データが挙げられる。他にも会社四季報のよ

うな企業単位のデータ、API サービス等で入手出来る Web 上の店舗情報など適用範囲は幅広い。

本システムは様々な地域や時点のデータに対応出来る必要がある。そこでシステム開発時のテストデータには住宅地図の入居者情報（『ZmapTOWN II』・株式会社ゼンリン）と電話帳データ（『タウンページデータベース』・NTT 情報開発株式会社）を用いた。日本のほぼ全域を同一の精度で途切れなくカバー出来ることからテストデータとして最適と判断した。

なお本論文はこれまで段階的に筆者らが行ってきた手法開発・改良を整理し、全国データに適用した結果を述べたものとなっている。

2. 開発

ソースデータの種類によってそれらが保有する空間情報の詳細さは異なる。そこで本システムでは空間情報の名称情報（施設名称、店舗名称等）と位置情報（住所、経緯度、建物情報）のみを使って異なる種類のデータ同士の空間結合および、異なる時点間の同一性判定を実現する。最終的に個店単位のデータ結合や、二時点間の時系列変化情報（存続、入替、新規出現、消滅）の取得が可能となる。

2.1. 開発環境と言語

市販されている一般的なパソコンを用いて開発した（Intel Core2 Quad Q6700 2.66GHz、メモリ 3.23GB、OS: Windows XP SP3）。開発言語は Ruby 1.9.1 および Ruby 1.8.6 である。一般的なパソコンと環境で開発を行った理由は、将来的なシステム公開に向けて本システムが特別な環境下のパソコンで無くとも動かすことが出来るよう考慮したためである。

2.2. ソースデータ

入力するソースデータはコンマ区切りのテキストデータ（csv 形式等）とする。名称情報と位置情報をコンマ区切りで入力したデータであれば処理可能である。位置情報は少なくとも経緯度を含む必要がある。住所、建物情報（建物名・階・部屋番号等）を追加することで更に高精度な空間結合が可能になる。また建物情報は建物名、階、部屋番号の分かち書き

統合元データ					統合先データ				
名称	位置情報				名称	位置情報			
	住所	建物情報	経度	緯度		住所	建物情報	経度	緯度
マクドナルド	東京都世田谷区	下北ビル301	139.66	35.66	マクドナルド下北沢店	東京都世田谷区	下北ビル3階	139.66	35.66
セタガヤ治療院	東京都世田谷区	下北ビル2F	139.66	35.66	鈴木自転車	東京都世田谷区		139.69	35.63
佐藤クリニック	東京都世田谷区	ハイツ佐藤1F	139.67	35.67	セタガヤ治療室	東京都世田谷区	下北ビル2階	139.66	35.66
鈴木自転車	東京都世田谷区		139.69	35.63	田中洋品店	東京都世田谷区	田中ビル1階A号	139.63	35.61
ロッテリア下北沢店	東京都世田谷区	下北ビル1F	139.66	35.66	佐藤歯科医院	東京都世田谷区		139.67	35.67
中華料理北京	東京都世田谷区	山田ヒルズ	139.65	35.65	中華上海	東京都世田谷区	山田ヒルズ	139.65	35.65

空間結合結果									
統合元名称	統合先名称	位置情報						結果	
		住所	建物情報	階	部屋番号	経度	緯度		
マクドナルド	マクドナルド下北沢店	東京都世田谷区	下北ビル	3	301	139.66	35.66	統合成功	
セタガヤ治療院	セタガヤ治療室	東京都世田谷区	下北ビル	2		139.66	35.66	統合成功	
佐藤クリニック	佐藤歯科医院	東京都世田谷区	ハイツ佐藤	1		139.67	35.67	統合成功	
鈴木自転車	鈴木自転車	東京都世田谷区				139.69	35.63	統合成功	
ロッテリア下北沢店		東京都世田谷区	下北ビル	1		139.66	35.66	統合元のみ	
	田中洋品店	東京都世田谷区	田中ビル	1	A	139.63	35.61	統合先のみ	
中華料理北京	中華上海	東京都世田谷区	山田ヒルズ			139.65	35.65	名称不一致	

注：上記の店舗・施設・建物等の名称、住所、経緯度等は全て架空のものである。

図1 ソースデータとデータ結合の出力結果のイメージ

新しい時点のデータ					古い時点のデータ				
名称	位置情報				名称	位置情報			
	住所	建物情報	経度	緯度		住所	建物情報	経度	緯度
ロッテリア大通店	北海道札幌市	大通ビル1階	141.35	43.06	ロッテリア札幌大通店	北海道札幌市	大通ビル1F	141.35	43.06
養老の瀧	北海道札幌市	大通ビルB1階	141.35	43.06	つば八大通り店	北海道札幌市	大通ビルB1F	141.35	43.06
札幌観光(株)	北海道札幌市	大通ビル303号	141.35	43.06	北洋商事	北海道札幌市	大通ビル202	141.35	43.06
北海道食品	北海道札幌市	鉄北会館2A	141.34	43.07	北海道食品(株)	北海道札幌市	鉄北会館3B号	141.34	43.07
札幌かに道場	北海道札幌市	薄野タワー202	141.35	43.05	かに道場	北海道札幌市	薄野タワー2階	141.35	43.05
セイコーマート琴似	北海道札幌市	ハイム琴似1F	141.31	43.07	セイコーマート	北海道札幌市		141.31	43.07

時空間結合結果									
新名称	旧名称	位置情報						結果	
		住所	建物情報	階	部屋番号	経度	緯度		
ロッテリア大通店	ロッテリア札幌大通店	北海道札幌市	大通ビル	1		141.35	43.06	存続	
養老の瀧	つば八大通り店	北海道札幌市	大通ビル	-1		141.35	43.06	入替	
札幌観光(株)		北海道札幌市	大通ビル	3	303	141.35	43.06	新規出現	
	北洋商事	北海道札幌市	大通ビル	2	202	141.35	43.06	消滅	
北海道食品		北海道札幌市	鉄北会館	2	A	141.34	43.07	新規出現・他の階から移入	
	北海道食品(株)	北海道札幌市	鉄北会館	3	B	141.34	43.07		
札幌かに道場	かに道場	北海道札幌市	薄野タワー	2	202	141.35	43.05	存続・階情報追加	
セイコーマート琴似	セイコーマート	北海道札幌市	ハイム琴似	1		141.31	43.07	存続・建物情報追加	

注：上記の店舗・施設・建物等の名称、住所、経緯度等は全て架空のものである。

図2 新旧2時点のデータの時空間結合（時系列化）の出力結果のイメージ

がされていないデータでも処理可能とする。図1、図2にソースデータと出力結果のイメージを示す。

2.3. 空間結合技術

一般的なGISソフトウェアの空間結合では空間データの二次元的な空間情報しか参照出来ない。住所表記の不完全さ等によるアドレスマッチング精度の低下は直ちに空間結合精度の低下に繋がる。また同一地点に複数のデータが存在する場合（高層ビルの入居テナントデータ等）結合先の特定が出来ない。

そこで本システムでは住所や建物情報（入居する階や部屋番号）も参照出来る三次元空間結合を実現

することでこの課題に対処した。図3に空間結合のアルゴリズムを示す。本処理は空間的に最も類似した位置情報を持つデータ同士を結合することに等しく、人間が目視で結合を行う過程を可能な限り忠実に再現したものである。候補が複数存在する場合は、名称情報が最も類似するデータと結合される。名称情報の類似度算出方法は次節で紹介する。

建物情報は1つの文字列で表現される場合が多い。即ち建物名称、階、部屋番号の分かち書きがされない。この文字列から建物名、階、部屋番号を認識させなければ、三次元的な位置情報は取得出来ない。そこで2005年の南関東全域の住宅地図の入居者情

報と電話帳データが持つ建物情報を用いて、建物情報の分割パターンライブラリを作成した。このライブラリを用いることで、建物情報から建物名称、階、部屋番号を認識させることが可能となる。図4に建物情報分割処理の例を示す。また建物情報分割処理のパターンライブラリの一部を表1に示す。建物情報の文字列内に表1に示す文字列が出現した場合、表1のルールに従って階と部屋番号を認識出来る。

本研究で開発した位置情報の結合技術は経緯度だけでなく、住所や建物情報といった非定量的な位置情報の処理も可能な点が優れていると言える。

図3 空間結合のアルゴリズム

図4 建物情報分割処理の例

2. 4. 結合後データの同一性判定

結合後データ同士の同一性の判定を行い、それらが同じデータであるか否かを決定する。すなわち空間的には結合されていても中身が異なるケースの発見を行う。また時系列化の場合は結合後データの同一性判定を行うことで二時点間の時系列変化情報を

表1 建物情報分割パターンライブラリ (一部)

パターン	階	部屋番号
[1-9][階]Fカイ	[1-9]	
[1-9][0-9][階]Fカイ	[1-9][0-9]	
[地下]B[チカ][1-9][階]Fカイ	[1-9]	
[地下]B[チカ][1-9]	[1-9]	
[1-9][階]Fカイ[0-9][0-9][号]ゴウ	[1-9]	[0-9][0-9][0-9]
[1-9][階]Fカイ[0-9][0-9][号]ゴウ	[1-9]	[0-9][0-9]
[1-9][階]Fカイ[0-9][号]ゴウ	[1-9]	[0-9]
[1-9][階]Fカイ[A-Z][号]ゴウ	[1-9]	[A-Z]
[1-9][階]Fカイ[A-Z]	[1-9]	[A-Z]
[1-9][0-9][階]Fカイ[0-9][0-9][0-9][号]ゴウ	[1-9][0-9]	[0-9][0-9][0-9]
[1-9][0-9][階]Fカイ[0-9][0-9][号]ゴウ	[1-9][0-9]	[0-9][0-9]
[1-9][0-9][階]Fカイ[0-9][号]ゴウ	[1-9][0-9]	[0-9]
[1-9][0-9][階]Fカイ[A-Z][号]ゴウ	[1-9][0-9]	[A-Z]
[1-9][0-9][階]Fカイ[A-Z]	[1-9][0-9]	[A-Z]
[地下]B[チカ][1-9][階]Fカイ[0-9][0-9][号]ゴウ	[1-9]	[0-9][0-9][0-9]
[地下]B[チカ][1-9][階]Fカイ[0-9][0-9][号]ゴウ	[1-9]	[0-9][0-9]
[1-9][0-9][0-9][号]ゴウ	※1	[1-9][0-9][0-9]
[1-9][0-9][0-9]	※1	[1-9][0-9][0-9]
[1-9][0-9][0-9][号]ゴウ	※2	[1-9][0-9][0-9][0-9]
[1-9][0-9][0-9]	※2	[1-9][0-9][0-9][0-9]
[0-9][0-9][号]ゴウ		[0-9][0-9]
[0-9][A-Z][号]ゴウ		[0-9][A-Z]
[1-9][階]Fカイ[0-9][号]ゴウ	[1-9]	[0-9]
[地下]B[チカ][1-9][階]Fカイ[0-9][号]ゴウ	[1-9]	[0-9]
[1-9][階]Fカイ[A-Z][号]ゴウ	[1-9]	[A-Z]
[地下]B[チカ][1-9][階]Fカイ[A-Z][号]ゴウ	[1-9]	[A-Z]
[0-9][号]ゴウ		[0-9]
[A-Z][号]ゴウ		[A-Z]
[地下]チカ\$	地下	
[A-Z][棟]トウ[0-9][0-9][0-9][号]ゴウ	半地下	
[A-Z][棟]トウ[0-9][0-9][号]ゴウ	数字一桁目	[0-9][0-9][0-9]
[A-Z][棟]トウ[0-9][号]ゴウ		[0-9]
管理人室		管理人室

注) [1-9]は1,2,3,4,5,6,7,8,9のいずれかにマッチングした場合という意味
[A-Z]はアルファベットのいずれかにマッチングした場合という意味
[階]Fカイは階、F、カイいずれかにマッチングした場合という意味
英数字は半角全角ともに認識可能
\$は語尾という意味
※1: [1-9][0-9][0-9]の3桁目を階にする。 例) 204号室 → 階: 2階 部屋番号: 204号室
※2: [1-9][0-9][0-9][0-9]の3, 4桁目を階にする。 例) 1404号室 → 階: 14階 部屋番号: 1404号室

統合元データ	統合先データ	統合先有無	名称同一性	結果
A	A	有り	一致	統合成功
B	C	有り	不一致	名称不一致
D		無し(統合元のみ)	---	統合元のみ
	E	無し(統合先のみ)	---	統合先のみ

図5 データ結合で取得出来る結果

新しいデータ	古いデータ	統合先有無	名称同一性	結果
A	A	有り	一致	存続
B	C	有り	不一致	入替
D		無し(統合元のみ)	---	新規出現
	E	無し(統合先のみ)	---	消滅

図6 時空間結合で取得出来る結果

取得出来る。同一性判定には双方が保有する名称情報を用いる。図5にデータ結合で取得出来る結果、図6に時空間結合で取得出来る結果を示す。

名称の同一性の判定は容易ではない。双方の名称が完全一致か否かという単純な判定結果では適切な結果が得られないためである。例えば同じ店舗や事業所を意味していても、データ間や時点間でその表記に揺れが生じる。表2に表記揺れの例を示す。表2は2005年における小田急線経堂駅周辺の住宅地図の入居者情報と電話帳データを手作業にて結合した結果の一部である。表2で示した30件のうち、名称情報が完全一致するデータは6件しか存在しない。

このような課題の解決には単語や文字列同士の同一性を定量化する必要がある。

そこで本研究では n-gram を用いて名称の類似度定量化を行った。n-gram は自然言語処理の一手法であり、異なる2つの文字列や文章の類似性を明らかに出来る。文学や言語学等の分野でも注目されている手法である（近藤，2000；村田ほか，2002）。

本研究では隣り合う2文字を次々と取り出して同一性を比較する bi-gram (2-gram) を用いる。本手法を用いることで表記揺れの問題を解決出来る。図7に bi-gram による処理の例を示す。文字列 i と文字列 j の類似度は式1によって定義される。

$$S_{ij}^{(n)} = \frac{n_{ij}^{(n)} + n_{ji}^{(n)}}{m_i^{(n)} + m_j^{(n)}} \quad (1)$$

$S_{ij}^{(n)}$ ：文字列 i と文字列 j の類似度

$m_i^{(n)}$ ：文字列 i から作成した n 文字の文字群の数

$m_j^{(n)}$ ：文字列 j から作成した n 文字の文字群の数

$n_{ij}^{(n)}, n_{ji}^{(n)}$ ： $m_i^{(n)}$ と同じ $m_j^{(n)}$ の個数

本手法による文字列の同一性判定では $S_{ij}^{(n)}$ の閾値を決定する必要がある。そこで2005年の東京都の電話帳データから3,000件を無作為に抽出し、それらのデータを同一地点にある2000年の電話帳データと結合し、結合したデータ同士の名称の類似度を上記の手法にて算出した。同時にそれらの一致不一致を目視にて確認することで $S_{ij}^{(n)}$ の閾値を決定した。図8にその結果を示す。本システムは閾値を任意に設定出来るが、図8の結果より $S_{ij}^{(n)}$ の閾値は0.4付近が適切であることが分かる。 $S_{ij}^{(n)}$ が0.4以上の場合、そのデータ同士を一致と見なす。以上より本システムでは $S_{ij}^{(n)}$ の値をデフォルトで0.4と設定した。

2.5. 雑音語の除去

店舗や施設等の名称には頻出語、地名、駅名が含まれる場合がある。これらの影響で前述した n-gram のみに頼る名称の同一性判定では適切な結果が得られない。相良ほか（2006）も純粋な店舗や施設の名称の抽出の難しさを指摘している。名称情報から頻出語や地名、駅名等（本研究では「雑音語」と呼ぶ）を除去する必要がある。

表2 データ間の表記揺れの例

住宅地図入居者データの名称情報	電話帳データの名称情報	完全一致
医療法人社団五十畑会五十畑クリニック	五十畑クリニック	
(有)スピリットオブハート	有限会社スピリット・オブ・ハート	
(株)柴田設計	株式会社柴田設計	
スキップフィール	スキップフィール	○
セブンイレブン世田谷経堂1丁目店	セブンイレブン 世田谷経堂1丁目店	
アイデント経堂教室	アイデント 経堂教室	
農大通り接骨院	農大通り接骨院	○
経堂ベッククリニック	経堂ベッククリニック	○
カラオケすずきフロント	カラオケすずき	
どんかつ新宿さぼてん経堂店	さぼてん デリカ経堂店	
オリジン弁当経堂城山通り店	オリジン弁当経堂城山通り店	
牛角経堂店	牛角 経堂店	
(株)三共商事	株式会社三共商事	
フォトショップあるばむ経堂店	フォトショップあるばむ 経堂店	
モスバーガー 経堂店	モスバーガー 経堂店	
ハンディクラフト	ハンディクラフト	○
財団東京勤労者医療会農大通り診療所	農大通り診療所	
オンリーワン	ライフハウス・オンリーワン	
くいもの処百姓亭	百姓亭	
本のヒロタ	本のヒロタ経堂店	
バブツイギー	バブツイギー	○
居酒屋らん	らん	
サンクス経堂農大通り店	サンクス経堂農大通り店	
かば田	株式会社かば田食品	
ローソン 経堂一丁目店	ローソン 経堂一丁目店	
ちよだ鮎	ちよだ鮎経堂店

図7 bi-gram による名称の類似度算出

図8 一致と不一致の bi-gram の件数分布

そこで頻出語、地名、駅名の辞書を作成し、それらと照合し雑音語を除去することでこの問題に対応した。頻出語は2000年と2005年の東京都全域の住宅地図の入居者名称（個人宅を除く）と電話帳データのテナント名称を形態素解析することで作成した。形態素解析には形態素解析システム『茶釜』を用いた。表3に形態素解析の結果得られた頻出語の一部を示す。地名は『日本行政区画便覧データファイル（日本加除出版株式会社）』から、駅名は現在および過去の時刻

表から駅名を収集し作成した。地名辞書には全国の丁目スケールまでの全地名、駅名辞書には1945年以降の全国全ての鉄道駅名を収録している。表4に雑音語の例を示す。また表5に各種雑音語辞書の単語登録件数を示す。なお地名や駅名を除去する際は図9のルールに当てはまる文字列を処理する。このルールの適用で地名や駅名を意味する文字列以外を除去するリスクを大幅に低減出来る。

2.6. ローカル頻出語の除去

前述した雑音語の除去だけでは不十分な場合がある。時間的、空間的に偏りが見られる頻出語が存在するためである。本研究ではこうした頻出語を「ローカル頻出語」と呼ぶ。

図10を用いてローカル頻出語を具体的に説明する。ローカル頻出語とは時空間的に出現に偏りが見られる頻出語である。「新宿西口」や「八重洲口」は地名ではない。また「八重洲口」は駅名でもない。しかしこれらの文字を含む店舗や事業所等は新宿駅西部や東京駅東部で数多く見られる。これらが「空間的に偏りがある」頻出語の例である。図11に2005年の電話帳データにおける「新宿西口店」をテナント名に含むデータの分布を示す。該当データが新宿駅西側に局所的に多数分布していることが分かる。

一方「六本木ヒルズ」は空間だけでなく時間的にも偏りが見られる頻出語の例である。六本木ヒルズの開業は2003年であり、2003年以前の店舗や事業所名が「六本木ヒルズ」という文字列を含むことは殆ど無い。例えば東京都の電話帳データの場合、「六本木ヒルズ」をテナント名称に含むデータは2005年の場合で105件、2009年の場合は141件存在するが、2000年のデータには0件である。このようなローカル頻出語を除去する技術が必要である。

そこで本研究ではソースデータの経緯度を用いて経緯度1/1000度四方のグリッドを作成し、グリッドごとに頻出する文字列を検索する。頻出する文字列の検索にはn-gramを用いる。グリッドごとにそのグリッドと隣接するグリッドを含めた全9グリッドに含まれる全データの名称情報を用いて9-gramから4-gramを作成する。周囲のグリッドを含む理由

表3 2005年における頻出語の出現回数

住宅地図件数(2005年): 3141434
電話帳件数(2005年): 690183
総件数: 3831617

頻出語	出現回数	出現率(%)※1
(株)	532007	13.88
(有)	322517	8.42
株式会社	151421	3.95
有限会社	69205	1.81
センター	54179	1.41
(事)	39617	1.03
美容室	32534	0.85
(営)	28489	0.74
ビル	27510	0.72
クリニック	19028	0.50

※1: 出現率=出現回数/総件数

注)実際の頻出語辞書には963種類の単語が登録されている。

表4 名称情報内に見られる雑音語の例

名称情報	雑音語	種類
グレイスジャパン株式会社	株式会社	頻出語
サーティワンアイスクリーム 麻布店	麻布店	地名
てもみん 新橋駅前店	新橋駅前店	駅名
まんが喫茶 グラグラ赤坂店	まんが喫茶	頻出語
	赤坂店	地名
アコム株式会社 新橋駅前支店	株式会社	頻出語
	新橋駅前支店	駅名
株式会社アイシービーアイワイジャパン白台店	株式会社	頻出語
	白台店	地名

表5 各種雑音語辞書の単語登録件数

地方(都道府県)	収録件数		
	地名辞書	駅名辞書	頻出語
北海道地方(北海道)	48570	1519	963
東北地方(青森・岩手・秋田・宮城・山形・福島)	135436	1314	
北関東地方(茨城・栃木・群馬)	12224	678	
南関東地方(埼玉・千葉・東京・神奈川)	35869	2625	
甲信越地方(新潟・長野・山梨)	16677	785	
北陸地方(富山・石川・福井)	14248	721	
東海地方(静岡・愛知・岐阜・三重)	76361	2008	
近畿地方(滋賀・京都・奈良・和歌山・大阪・兵庫)	66681	2290	
中国地方(鳥取・島根・岡山・広島・山口)	21554	1081	
四国地方(香川・徳島・高知・愛媛)	17303	580	
九州地方(福岡・長崎・佐賀・大分・熊本・宮崎・鹿児島)	28868	1717	
沖縄地方(沖縄)	1835	19

図9 削除対象となる地名・駅名を含む文字列

はグリッドの境界付近のデータにも適切な処理を行うためである。除去対象となる文字列はそれらのうち複数回の出現が見られ、かつ語尾に「店」、「支社」、「営業所」、「東口」、「西口」、「南口」、「北口」を含むものとする。また除去の際には可能な限り長い文字列を確実に除去するために文字数の多い頻出語から順に除去していく。すなわち9-gramで得られた頻出語から除去し、最後に4-gramで得られたもの

を除去する。図12は新宿駅西側のあるグリッドに見られるローカル頻出語とその除去処理の例である。一部で雑音語が残る場合や余分な除去を行ってしまう例も見られるが、多くは適切に処理されている。なお雑音語が部分的に残ったり、余分な削除を行ったりした場合でも、名称の類似度判定に n-gram を用いるため、その影響をほぼ無視出来る。

表6、表7に本システムによる各種雑音語除去の成果を示す。電話帳データの場合(表6)、全ての雑音語除去のための処理を行うことで、目視で雑音語除去が必要と判断された654件のうち、92.4%が店舗・事業所名を壊すことなく雑音語のみを取り除くことに成功した。表記ゆれが大きいWebから得られる情報(表7)の場合でも、79.4%のデータで雑音語のみを取り除くことが出来た。また何れの処理の場合も頻出語辞書による効果が最も大きかった。

図10 ローカル頻出語のイメージ

GridID:13970035686におけるローカル頻出語一覧

n	ローカル頻出語	出現回数	削除前テナント名	削除後テナント名	例
9	新宿西口国際通り店	2	福膳新宿西口国際通り店	福膳	
8	新宿エステック店	3	ちよだ館新宿エステック店	ちよだ館	
8	エステック情報店	3	そじ坊 新宿エステック情報店	そじ坊	
8	ク 新宿南口支店	2	ディック 新宿南口支店	ディック	
7	ト 新宿南口店	2	協和コンタクト 新宿南口店	共和コンタクト	
6	新宿南口店	3	北前そば高田屋 新宿南口店	北前そば高田屋	
6	新宿駅西口店	2	ミニストップ新宿駅西口店	ミニストップ	
5	新宿南口	8	天狗 新宿南口	天狗	
5	新宿西口	4	ボーダフォン 新宿西口	ボーダフォン	
5	新宿南口店	3	山頭火新宿南口店	山頭火	
5	東京支社	2	日本インター 東京支社	日本インター	
4	新宿西口	8	キリンシティ新宿西口	キリンシティ	
4	新宿南口	4	リテラ・クレア リテラ新宿南口	リテラ・クレア リテラ	
4	新宿支社	3	ジャパンプロテック新宿支社	ジャパンプロテック	
4	西口本店	2	ヨドバシカメラ 西口本店	ヨドバシカメラ	
4	2号店	2	マップカメラ 2号店	マップカメラ	
4	東京本社	2	近畿設備東京本社	近畿設備	
4	西口店	2	ウスキュダル 西口店	ウスキュダル	

※検索範囲は南西端N35.685 E139.699、北東端N35.688 E139.702の3/1000° 四方。
※削除前テナント名はローカル頻出語以外の頻出語はクリーニング済み。

図12 新宿駅西側の1グリッドにおけるローカル頻出語とその除去の例(2005年・電話帳データ)

図11 「新宿西口店」をテナント名に含むデータの分布(2005年・電話帳データの場合)

表6 本システムによる各種雑音語除去の結果を手作業で得られた結果と比較した成果(その1)

雑音語除去の成果(電話帳の場合:東京都全域から1000件をランダム抽出)

総件数	1000			
手作業にて確認した場合に雑音語除去が必要と確認されたか?	○:654		×:346	
頻出語辞書による雑音語除去で手作業と同じ結果を得られたか?	○:513	×:141		
さらに地名・駅名辞書による雑音語除去で手作業と同じ結果を得られたか?	○:70	×:71		
さらにローカル頻出語雑音語除去で手作業と同じ結果を得られたか?		○:11	×:60	
上記の雑音語処理後も処理前の文字列が保たれているか?			○:330	×:16
正しく処理された件数	513	70	11	330
				924
				処理精度(%) 92.40

表7 本システムによる各種雑音語除去の結果を手作業で得られた結果と比較した成果(その2)

雑音語除去の成果(hotpepper APIで取得した店舗名:東京都全域から1000件をランダム抽出)

総件数	1000			
手作業にて確認した場合に雑音語除去が必要と確認されたか?	○:545		×:455	
頻出語辞書による雑音語除去で手作業と同じ結果を得られたか?	○:87	×:478		
さらに地名辞書による雑音語除去で手作業と同じ結果を得られたか?	○:169	×:309		
さらに駅名辞書による雑音語除去で手作業と同じ結果を得られたか?	○:88	×:241		
さらにローカル頻出語雑音語除去で手作業と同じ結果を得られたか?		○:81	×:160	
上記の雑音語処理後も処理前の文字列が保たれているか?			○:409	×:46
正しく処理された件数	87	169	88	409
				794
				処理精度(%) 79.40

2.7. 処理の高速化

空間検索を行う際に入力データが膨大な場合、問い合わせ先データ件数も膨大な数となる。そこで経緯度情報を用いて、あらかじめ入力データにそれらが持つ経緯度に基づいたIDを付加し、経緯度1/1000度四方のグリッドに分配する。空間検索時には問い合わせ元データ自身が含まれるグリッド内での位置に応じて、検索範囲を可変させることで処理の大幅な高速化を実現した。図13に本研究の空間インデックスの手法と検索範囲決定のルールを示す。また表8に空間インデックスを導入することによる計算速度の改善状況を示す。データ量が多い上に、狭い範囲に多数のデータが存在しデータ処理に時間がかかると考えられる都市部ほど本手法の効果が大きい。

3. 処理精度の検証

南関東地方の幾つかの地域をサンプルにして本手法にて自動的に作成した結果と手作業にて作成した結果（正解データ）を比較する。検証には住宅地図の入居者データと電話帳データを用いる。サンプル抽出地域を表9に示す。都市部のサンプルとして繁華街と駅前商店街の地域を用いた。また地方都市とデータの分布が粗な地域として離島の集落を用いた。

まず2000年と2005年の南関東全域の全データ（住宅地図と電話帳の結合はともに2005年）を本システムによって結合する。その後対象地域の全件を地域ごとに手作業で作成した結果と比較する。

表10、表11、表12にシステムの処理精度を示す。何れの表もサンプルの4地域の結果の総計である。住宅地図同士の時空間結合が94.36%(820/869)、電話帳同士の時空間結合が95.22%(478/502)、両データ結合が93.77%(767/818)の精度で処理出来ることが明らかになった。何れの結果もその総件数が表9の総件数と一致しない理由は、消滅（結合の場合は住宅地図のみ）が出力されるためである（例えば表10の消滅を除いた件数の和は2005年の住宅地図の総件数に一致する）。

また表13にはサンプル地域ごとの住宅地図同士の時系列結合の処理精度を示す。例えば歌舞伎町の存続「79(87)」とは手作業で処理した結果87件が存続

と判定され、そのうち79件を本システムも存続と判定した、という意味である。処理精度は各時系列変化の総数のうち、本システムが手作業と同じ結果と判定した件数の割合を示す。都市部では地方都市や離島よりも処理精度がやや低下した。また同じ都市部同士でも建物情報が豊富かつ正確に収録されていると考えられる歌舞伎町は処理精度が高かった。地方都市や離島ではほぼ100%の処理精度となった。

なお手作業で同様のデータ作成作業を行っても5%前後の作業ミスが発生することが分かっている。処理精度検証のための正解データ作成の際にも本システムで得られた結果を用いることで、手作業による作業ミスを効率よく発見出来ている。このように手作業でもヒューマンエラーが発生する上に、その作業量、作業時間が膨大なものになることを考慮すると、本システムの処理精度は充分に実用に堪え得るものであると言える。

図13 空間インデックス

表8 空間インデックス有無による計算時間の差

入力データ	入力データ(新)	タウンページデータベース(2005年)世田谷区全域:28805件
処理方法	入力データ(古)	タウンページデータベース(2000年)世田谷区全域:32224件
		時系列化
PCスペック	OS	Windows XP Professional Service Pack 3
	CPU	Intel® Core™ 2 Quad CPU Q6700 2.66GHz
	メモリ	3.23GB
計算時間	空間インデックス有り(T1)	891.89秒
	空間インデックス無し(T2)	207552.84秒
	T2/T1	232.71

表9 処理精度検証のためのサンプル地域

サンプル地域		件数(2005年)		件数(2000年)	
住所	地域特性	住宅地図	電話帳	住宅地図	電話帳
東京都新宿区歌舞伎町1丁目17・23番地	大都市の繁華街	191	88	210	94
東京都世田谷区経堂農大通り商店街	駅前商店街	276	130	247	141
千葉県鴨川市天津漁港周辺	港町	121	82	136	94
東京都八丈島八丈町中之郷・榎立地区	離島の集落	140	105	144	121
合計		728	405	737	450

表10 住宅地図の時系列化処理の処理精度

		システム						処理精度(%)
		総件数	存続	入替	新規	消滅	SIミス	
手作業	存続	467	443	4	22	0	2	94.86
	入替	129	4	121	8	0	0	93.80
	新規	132	4	2	126	0	0	95.45
	消滅	141	8	3	0	130	0	92.20
		総計						94.36

表11 電話帳の時系列化処理の処理精度

		システム						処理精度(%)
		総件数	存続	入替	新規	消滅	SIミス	
手作業	存続	307	293	4	10	0	0	95.44
	入替	46	1	43	2	0	0	93.48
	新規	52	0	3	49	0	0	94.23
	消滅	97	2	2	0	93	0	95.88
		総計						95.22

表12 住宅地図と電話帳の空間結合の処理精度

		システム						処理精度(%)
		総件数	統合成功	名称不一致	電話帳のみ	住宅地図のみ	SIミス	
手作業	統合成功	269	246	16	7	0	0	91.45
	名称不一致	46	4	42	0	0	0	91.30
	電話帳のみ	90	5	1	83	0	1	92.22
	住宅地図のみ	413	11	6	0	396	1	95.88
		総計						93.77

表13 サンプル地域ごとの処理精度比較

サンプル地域	2005年 総件数	2000年 総件数	住宅地図時系列処理精度(2000～2005年)					処理精度 (%)
			存続	入替	新規	消滅	SIミス	
東京都新宿区歌舞伎町1丁目17・23番地	191	210	79 (87)	67 (74)	28 (30)	44 (49)	2	94.58
東京都世田谷区経堂農大通り商店街	276	247	137 (154)	48 (51)	67 (71)	38 (42)	0	92.45
千葉県鴨川市天津漁港周辺	121	136	116 (116)	0 (1)	4 (4)	19 (19)	0	99.29
東京都八丈島八丈町中之郷・櫻立地区	140	144	106 (110)	2 (3)	27 (27)	29 (31)	0	97.66

4. 作成データの例

4.1. ミクロスケールデータ

図14は新宿駅周辺の2000～2005年のテナント交代の3Dマップである。2000年と2005年の電話帳データを時系列化した後に、住宅地図の入居者データと結合して作成した。視覚化の都合上、1フロアに1つのデータしか入っていない様に見えるが、実際のデータは部屋単位である。また全フロアが新規出現の建物を探すことで2時点間に新築された建物の発見を行ったり、全フロアが消滅の建物を探すことで2005年時点に再開発中や更地になっている場所を発見したりすることが可能となる。新宿駅周辺では低層階ほど店舗の入替が比較的多く起こっていることが分かった。また新宿御苑東側で新築建物が多数見られることも分かった。

4.2. マクロスケールデータ

日本全土の2003年頃と2008年頃の住宅地図の入居者データ（個人宅以外）を時系列化し、4次メッ

図14 新宿駅周辺のテナント交代3Dマップ

シュごとにメッシュ内の総件数に対する入替データの割合（入替率）を明らかにした（図15）。広域的に見ると店舗間競争が盛んであると考えられる都市部で高い値を示すことがよく分かる結果となった（図16）。また同じ都市の中でも場所によってその値に大きなばらつきが見られることは興味深い。例えば図16の京都市中心部の図では特に祇園四条周辺で非常に高い値を示しており、京都市内でも特に店舗間競争が激しい地域であることが分かった。

これほど広域を連続的に網羅した詳細な時空間データは前例が無い。本データは図14に見られる高精度な時系列データから作成されており、様々なスケールの時空間分析に利用可能である。このようなデータは1章で紹介した課題の解決に貢献出来るものである。

ただし現段階では南関東地方と南関東地方以外の一部地域でしか処理精度の検証が行われていないため、上記の全国データを用いて全国の複数箇所でも3章と同様の検証を進める必要があるだろう。なお既に札幌市と京都市の一部で表13と同様の検証が行われており（表14）、南関東地方と同程度の処理精度になることが確認されている。

4.3. Webデータとの結合

Hotpepper APIを用いて2009年10月頃に取得した『ホットペッパー』（株式会社リクルート）に掲載されたデータと住宅地図の入居者データ（2009年）を結合した例を図17に示す。Hotpepper APIを用いて取得した店舗データには店舗名、住所、経緯度、更に入居する建物の情報も収録されているため、本シ

図15 日本全土におけるテナント入替率マップ

図16 南関東地方(左上), 京都市中心部(右上)及び近畿中京圏(下)のテナント入替率マップ

表14 札幌市と京都市の一部における処理精度

サンプル地域	2005年 総件数	2000年 総件数	住宅地図時系列処理精度(2000~2005年)					処理精度 (%)
			存続	入替	新規	消滅	SIミス	
北海道札幌市北区北19条西5丁目周辺	249	219	148 (149)	8 (9)	87 (91)	57 (61)	0	96.77
京都府京都市東山区清水町周辺	205	216	100 (102)	72 (81)	20 (22)	31 (33)	0	93.70

※地域特性 札幌市: 中高層建物と低層建物が低密度に条里状に整然と分布する地域
京都市: 長屋状の低層建物が高密度に密集した地域

システムを用いることで両データをかなり正確に結合出来る。結合結果には店舗の情報だけでなく、店舗のURLも埋め込むことが出来るため、Web情報との連携も容易である。本システムを使えば図17の例に示したように異なる大規模な空間情報同士を正確かつ容易に結合することが可能になる。また1.2節で学術的な領域においてデータ作成時に多大な労力と時間を費やす場合があることを紹介したが、本システムはこれらの課題の解決に大いに貢献出来る。

図17 ホットペッパー掲載の店舗情報と住宅地図を結合しWebページを参照する様子

5. おわりに

本研究では異なる空間情報同士の結合及び異なる二時点間の空間情報の時系列化を実現するシステムの開発を行った。経緯度のみならず住所や建物情報といった非定量的な位置情報を用いた空間結合及び、曖昧さを持つテキスト同士の同一性判定技術を提案し実用化された。名称情報と経緯度を持つテキストデータであれば処理可能であり、建物情報が付加されたデータの場合は三次元空間結合に拡張出来る。

また本研究で開発した店舗等名称を用いた同一性判定や建物情報を利用した三次元空間結合の考え方は、これからの新しい高度なGISシステム開発のシーンにおいて貴重な知見となりうるだろう。例えば同一地点に複数の情報が分布するデータ(高層ビル・雑居ビル等の入居者情報)同士を正確に空間結合することは現状の一般的なGISソフトウェアでは実現されていないが、本研究の三次元空間結合の手法はその課題を解決出来る。

テキスト処理に空間情報(経緯度)を導入した手法(ローカル頻出語除去)は自然言語処理の領域において斬新な手法である。また本研究を通して今後は空間情報学や地理学の領域においても、増々の自然言語処理技術の導入が必要になる可能性があるという知見を得ることが出来た。

空間情報学や地理学、建築学などを中心とした学問的領域や、不動産・コンサルティング等の民間企業、また都市・地域計画等を行う国及び地方自治体等では、施設名称や店舗名称と住所情報を持つ名簿

状のデータ同士をその位置情報に応じて結合・突合して利用したい場合が多い。また結合後データの同一性を見極め、結合の成功可否を判定することも重要である。本研究で提案した手法は上記のような様々なシーンにおいて活用可能である。そのため今後は本システムやデータの利用環境の整備も重要な課題である。また多くの利用者に様々な地域や種類のデータを処理して頂くことで、本システムの改良点を明らかにし、更なる性能向上を目指したいと考えている。本研究を通して開発した住宅地図の入居者情報や電話帳データの時系列データセットについては現在、公開環境の整備を進めているところである。これらのデータが利用可能になれば、様々な研究領域において新しい基礎データの1つとして広く利用されるようになる可能性がある。なお既に上記のデータを用いて行われた研究も見られる (Osada et al., 2007; 澁木ほか, 2008; 関口, 2011)。

更に本手法で作成出来る時系列データは既存の統計データの代替データとしての活用可能性もある。既に我々はこれまでの研究で電話帳の時系列データを事業所・企業統計調査の代替補完のためのデータとして利用出来る可能性について検討を行い、その機能がある程度有していることを確認している (Akiyama and Shibasaki, 2009)。今後の研究で時系列データを既存統計情報の代替データとしての利用のための手法も明らかにしていく予定である。

謝辞

全国の住宅地図データの利用環境は株式会社ゼンリンとの共同研究によって実現した。ここに記して謝意を表したい。

参考文献

浅野純一郎 (2002) 地方都市中心市街地における大規模商業施設の閉店や郊外移転の実態とその後の利用・跡地利用の方向性～北陸甲信越地方の地方自治体担当部局への調査から～。「日本建築学会計画系論文集」, **577**, 257-264.

阿藤卓弥・大村謙二郎・有田智一・藤井さやか (2006) 首都圏郊外における鉄道駅前商業集積の停滞実態と

その課題: 本厚木駅・小田原駅前地区を対象として。「日本都市計画学会学術研究論文集」, **41**(3), 1037-1042.

伊藤香織・曲渕英邦 (2001) 既存情報を活用した時空間データ作成手法－地図内・地図外情報の曖昧性を考慮した空間要素同定を用いて－。「地理情報システム学会講演論文集」, **10**, 147-150.

近藤みゆき (2000) n グラム統計処理を用いた文字列分析による日本古典文学の研究－古今和歌集のことばの型と性差－。「千葉大学「人文研究」」, **29**, 187-238.

相良毅・牧野俊朗・川口修一・小澤英昭・喜連川優 (2006) 住所情報を用いた店舗名称のクリーニング手法－Web からの店舗情報収集精度向上のために－。「DEWS2006」, 2C-01, 392-398.

澁木猛・秋山祐樹・柴崎亮介 (2008) デジタル地図と電話帳データの時空間統合による店舗及び事業所の立体分布変動モニタリング手法。「日本建築学会計画系論文集」, **73**, 626, 789-793.

関口達也 (2011) 住宅地滲出型商業集積の形成過程とその展開－原宿・青山・代官山を事例とした時系列分析－。「東京大学大学院都市工学専攻平成22年度修士論文」

村田哲也・高橋勇・黒岩丈介・小高知宏・小倉久和 (2002) 学生レポートの n-gram による類似度評価の検討。「第1回情報科学技術フォーラム講演論文集」, E-10, 30.

Akiyama, Y., Shibasaki, R. (2009) Development of Detailed Spatio-temporal Urban Data through the Integration of Digital Maps and Yellow Page Data and Feasibility Study as Complementary Data for Existing Statistical Information. CUPUM2009, Reference No. 187 Session 8.2-2.

Osada, T., Akiyama, Y., Shibuki, T., Shibasaki, R. (2007) A Study on Distribution Tendencies of Business Establishments by Detailed Urban Dataset and Its Accuracy Validation. The 9th International Cooperative Seminar between KAGIS and GISA, 197.

(2011年1月31日原稿受理, 2011年7月11日採用決定, 2011年9月29日デジタルライブラリ掲載)