東京大学大学院新領域創成科学研究科環境学研究系
社会文化環境学専攻学位論文梗概集2006 pp.35-38

既存電子地図と電話帳情報の時空間的結合による詳細都市データセットの作成に関する研究

A Study on Efficient Making Method of Detailed Time-series Urban Dataset by Spatial Integration of Digital Maps and Yellow Page Data for Urban Analysis

学籍番号　56814
氏　　名　秋山 祐樹（Akiyama, yuki）
専　　攻　社会文化環境学専攻（2006 年度修了）
指導教員　柴崎 亮介 教授

ABSTRACT: There were many studies of urban analysis with various ways in various disciplines. But information used often has low spatial resolution especially when they use regional statistics, though the statistical data can over large areas with homogeneous quality. For detailed analysis, some studies rely on field survey that has very fine spatial resolution, but they fail to cover an entire urban area or large regions including suburb and rural area.
In this study, there are two aims. First is to develop a method of making time-series tenant dataset by integrating of multi-year yellow page. Second is to develop a method of generating a new digital map with time-series tenant data that can cover whole urban area or national land by integrating detailed digital maps with time-series tenant data.
All of this system is automated, so anyone can acquire results easily in short time with high accuracy. This dataset allows us to cover whole extent of Japan with “accuracy” and “flexibility”.

KEYWORD: Yellow page, Digital house map, Natural language processing, Time-series dataset

1. 詳細都市データセットの必要性

これまでに「都市」という空間の把握のために様々なデータセットが開発されてきている。しかしそれらは解像度や信頼性が高くてもごく限定的な地域を対象とした場合が多い。一方非常に広域をカバーしていたとしてもその解像度は低く、信頼性にばらつきのあるものが多い。現在日本各地の地方都市で起こっている都心部の衰退化や大規模小売店舗出店の影響等の分析には、複数都市を対象とした都市内部の状況を表す時系列データが必要と言われており [1)]、日本全土を対象とした高解像度でしかも信頼性にばらつきのない時系列化都市データセットを開発することは大変有意義であるといえる。

本研究では都市の変化を追うためにテナントの変遷に注目する。テナントの変遷は都市空間の時空間的変化を詳細に観察する場合に最適である。テナントの情報収集にはタウンページデータを用いる。複数年次のタウンページを組み合わせてテナントの時系列変化を追うことが出来るデータベースを作成する。更にタウンページと電子地図を結合することで詳細なテナント時系列情報と高い空間的精度を併せ持つ時空間地図を作成出来る。本研究ではこの時空間地図を「詳細都市データセット」と名付ける。

本研究の目標は詳細都市データセットの作成技術の開発である。将来的な最終目標は日本全土の詳細都市データセットの整備である。

2. データの性質

詳細都市データセットのソースデータには、テナントの時系列情報収集に「タウンページデータベース」、時系列情報をリンクさせる電子地図に電子住宅地図である「Zmap-TOWNⅡ」を用いる。何れの情報もテナント名称、入居する建物名称、階、部屋番号、住所を保有する。またタウンページはテナントの業種も保有する。一方 Zmap データは建物形状データを保有する。

3. 電話帳情報の時系列化

電話帳の時系列化は異なる年のデータ同士を位置情報と名称情報に基づいて正確に結合させる方法を開発する必要がある。

3-1 位置情報のリンク

電話帳の時系列化を行うにはまず異なる年のテナント同士の位置情報をリンクする必要がある。電話帳情報は住所と建物情報（入居する建物名称・階・部屋番号）を持つ。またアドレスマッチング[1]を利用することで緯度経度情報を付加することが出来る。Fig1 に位置情報リンクのアルゴリズムを示す。リンク元が持つ最良の条件の位置情報を認識し、その位置情報に最も近い場所とリンクさせる。

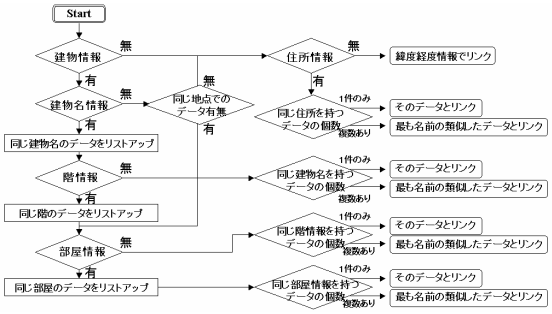

Fig1 位置情報リンクのアルゴリズム

建物情報は１つのテキストで、建物名称、階、部屋番号は分かれていない。そこで建物名称のパターンをライブラリ化し正規表現にて建物名称、階、部屋を分けて認識出来るようにした。Fig2 に建物情報の分析例を示す。

また最も名称の近いテナント検索には後述する名称情報比較アルゴリズムを用いる。

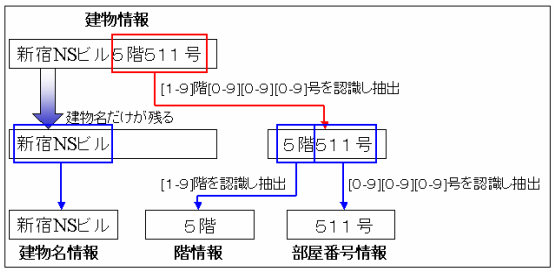

Fig2 建物情報認識アルゴリズム

3-2 名称情報の一致不一致判定

位置情報に基づいてリンクされたテナントの一致不一致の判定はテナント名称に基づいて行われる。すなわち名称の類似度合いを定量化する必要がある。テナント名称には同じテナントでもぶれ[2]があり完全一致では適切な判定は不可能である。本研究では名称類似度計算に自然言語処理方法の一つ「n-gram」を用いる。n-gram はテキストの類似度合いを数値化することが可能であり、近年では文学や言語学の領域でも注目されている[2),3),4)]。本研究では隣接する２文字を取り出して比較する bi-gram を主に用いる（Fig3）[5)]。

$m_i^{(n)}$と $n_{ij}^{(n)}$を以下のように定義する。

$m_i^{(n)}$：テキスト i から取り出した n 文字の組の個数

$n_{ij}^{(n)}$：$m_j^{(n)}$と一致する $m_i^{(n)}$の個数

テキスト i とテキスト j の類似度は Eq.1 となる。

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

一致不一致の閾値は 0.4 が適切であった。実際の計算では事前に頻出語や地名をクリーニングして更に精度を向上させている。

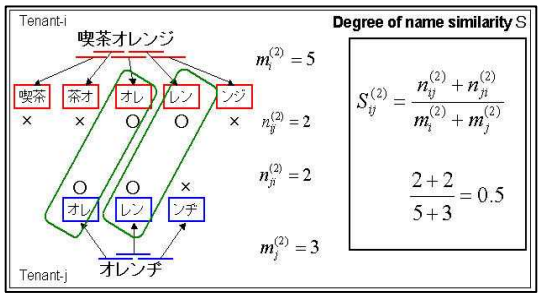

Fig3 n-gram（n=2 の場合）

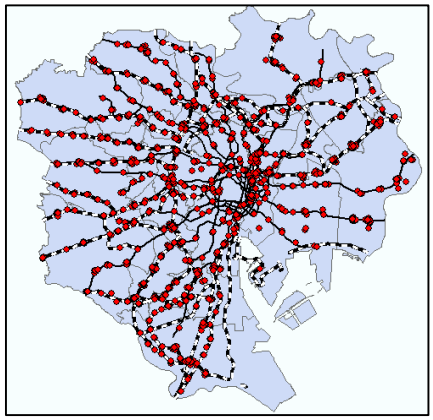

Fig4 ローカル頻出語の例：東京 23 区で「～駅前」を名称に含むテナントの位置

またある特定に地域に集中的に出現する「ローカル頻出語」(Fig4)のクリーニングシステムも搭載しており、更なる精度の向上が可能となった。

3-3 得られる時系列情報

Fig5 に本システムで得られる時系列変化の一覧を示す。このような時系列変化の情報が全てのテナントに付加され、過去のテナント名称や業種を参照可能になる。過去に消滅したテナントや、他フロアから移入してきたテナントの特定も可能である。Fig6 にテナント時系列情報を用いて作成したデータの例を示す。

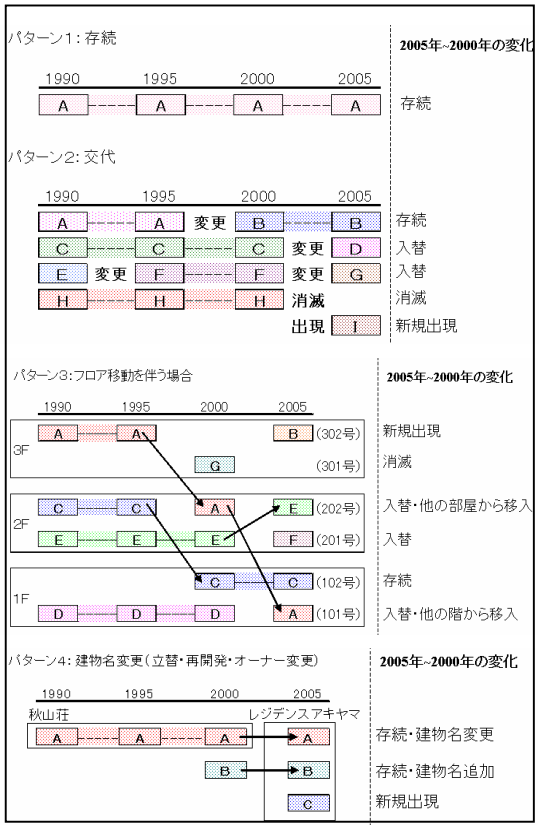

Fig5 特定可能な時系列情報一覧

4. 電話帳情報と電子住宅地図の結合

電話帳同士の時系列化のアルゴリズムの応用で可能である。共に位置、テナント名称、建物の情報を保有する。これらを基に電話帳同士の位置情報のリンクと同様の方法で正確な空間結合が可能である。名称が一致すれば結合成功である。調査年次のずれや調査精度によっては一致しない場合もある。その場合はエラーとなる。Fig7 に両データを結合して得られる３D マップの例を示す。

5. システムの処理精度

システムの処理精度を評価するために丁目単位で東京 23 区からサンプルを取り出し、手作業の結果と比較した[3]。Table1 に電話帳同士の時系列化の処理精度、Table2 に電話帳と電子住宅地図結合の処理精度を示す。入替、あるいは不一致の結果にやや不安が残るものの、それ以外はほぼ 100%近い処理精度であった。頻出語処理のアルゴリズムの改良や n-gram の閾値調整等で更なる改善が可能と考えられる。

Table1 電話帳時系列化の処理精度

		総件数	システムと一致	精度 (%)
手作業	存続	1004	975	97.11
	入替	121	104	85.95
	入替･移入あり	4	4	100.00
	新規	362	359	99.17
	新規･移入あり	8	3	37.50
	消滅	672	672	100.00

Table2 電話帳と住宅地図結合の処理精度

		総件数	システムと一致	精度 (%)
手作業	一致	510	479	93.92
	不一致	57	50	87.72
	Zmapのみ	684	674	98.54
	タウンのみ	188	188	100.00

6. 現地調査によるデータの信頼性評価

電話帳情報と住宅地図の信頼性を評価するために墨田区墨田４丁目および港区六本木４丁目にて現地調査を行い、ソースデータの信頼性を評価した。調査地区は本研究にて定義したテナント安定度[4]に基づき、安定度の高い地区と低い地区を選択した。Table3、Table4 に両地域のデータの信頼性を示す。調査に用いたデータと現地調査には約１年 6～10 ヶ月のタイムラグがあるにも関わらず良好な結果を得られた。特に電話帳と住宅地図が共に持つテナントは信頼性が高い。またテナントの動きが少ない墨田４丁目では現地調査で確認できるテナントを超える数のテナントを収集することが可能である。両データの結合によってお互いに不足したテナント情報を補完しあうことが出来るため、両データの結合はそれぞれ単体で利用するよりも収集できる情報を増やすことが出来る。

7. 結論と今後の展開

本システムによりテナントの時系列データセットの作成、そしてテナント時系列化データセットと電子住宅地図の結合が可能となった。しかも精度も高く処理時間も高速である。

今後はシステムの改良を進めると共に、現在我々が保有する南関東全域のデータを詳細都市データセットとして整備する。電話帳の時系列化についてはほぼ完了しており、今後は電子住宅地図との結合作業を行い、利用可能なデータセット整備を早急に行うことを目指す。

注釈

(1) CSV アドレスマッチングサービス

(2) 例えば渋谷区に 2000 年に存在した「ワイルドワン宇田川店」の 1995 年の名称は「ワイルドワンインターナショナル」。

(3) 伊藤 [5]が明らかにした東京のテナント時系列変化の傾向を用いて場所を選定。墨田区墨田４丁目、文京区根津２丁目、港区六本木４丁目、千代田区大手町１丁目、大田区田園調布３丁目。

(4) 存続テナント数を全テナント数で除した値。東京 23 区の全街区にて算出。

参考文献

1) 室町泰徳, 原田昇, 太田勝敏 (1994)「都心商業地域の衰退状況と大規模小売店舗の立地動向に関する研究」, 第 29 回日本都市計画学会学術研究論文集 29, pp.529-534

2) Masayuki Asahara, Yuji Matsumoto (2003) “Japanese Named Entity Extraction with Redundant Morphological Analysis” HLT-NAACL 2003, pp.8-15

3) 近藤泰弘, 近藤みゆき (2001)「平安時代古典文学研究のための N-gram を用いた解析方法」, 言語情報処理学会第 7 回年次大会『発表論文集』

4) 師茂樹 (2002)「N グラムモデルとクラスター分析を用いた漢文古典テキストの比較研究－『般若心経』の意訳の比較を例に－」, 京都大学大型計算機センター第 69 回研究セミナー「東洋学へのコンピュータ利用」予稿集

5) 伊藤香織 (2001)「都市空間の事象性に関する研究」東京大学大学院工学系研究科建築学専攻博士論文

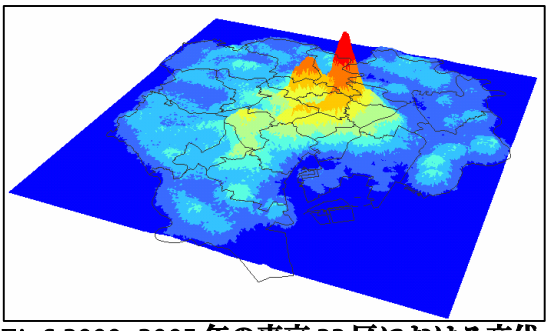

Fig6 2000~2005 年の東京 23 区における交代テナントの密度

Table3 港区六本木４丁目におけるデータの信頼性評価

件数			両方保有			電話帳のみ			Zmapのみ			収集可能件数							
Zmap	電話帳	実際	一致	不一致	信頼率	一致	不一致	信頼率	一致	不一致	信頼率	一致	不一致	総数	補完率	信頼率	調査不能	差し引き	信頼率
883	511	1046	310	68	82.01	51	53	49.04	318	152	67.66	679	273	952	91.01	64.91	49	903	75.19

Table4 墨田区墨田４丁目におけるデータの信頼性評価

件数			両方保有			電話帳のみ			Zmapのみ			収集可能件数							
Zmap	電話帳	実際	一致	不一致	信頼率	一致	不一致	信頼率	一致	不一致	信頼率	一致	不一致	総数	補完率	信頼率	調査不能	差し引き	信頼率
368	244	371	153	5	95.63	41	22	50.00	154	25	74.04	348	52	400	107.82	87.00	50	350	99.43

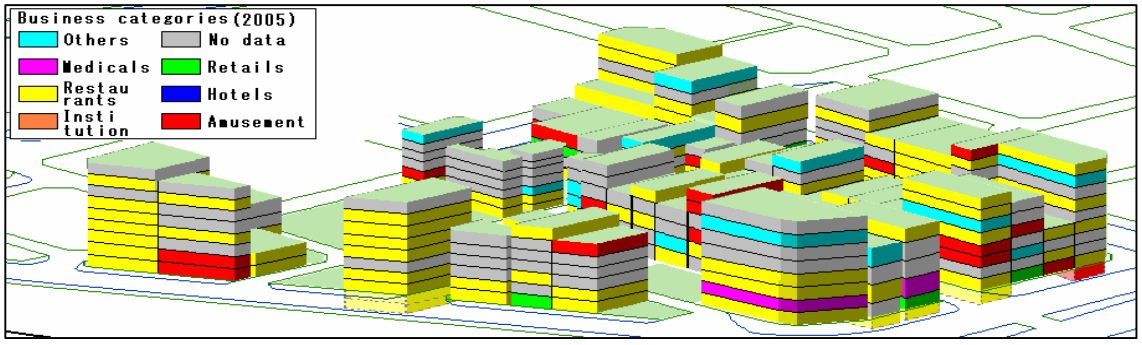

Fig7 新宿区歌舞伎町の業種分類３D 地図の一部（2005 年）