1D1-2

Web情報を利用した駅クラスタリング手法の提案

Clustring Stations based on Web Information

清　雄一　　小池　亜弥　　白井　康之
Yuichi Sei　　Aya Koike　　Yasuyuki Shirai

（株）三菱総合研究所
Mitsubishi Research Institute, inc.

ユーザの位置情報に基づいてサービスを提供するロケーションベースサービスが普及しつつある。ユーザの正確な位置情報が悪意のある第三者に渡るとユーザの居住地や職場が特定されてしまう脅威が存在するため、位置情報を含めたユーザの個人情報を保護できる能力を持ったサービス提供者でなければ、ロケーションベースサービスの提供はできない。だが、このような能力を持った事業者がユーザの位置情報を適切に管理し、ユーザを一意に特定できないよう匿名化された位置情報を作成することができれば、より多くの事業者がロケーションサービスの分野に参入できると考えられる。このように、ユーザを一意に特定できないように匿名化を行う指標の一つとして、k-匿名性が提案されている。これは、ユーザの位置情報を曖昧化し、曖昧化された領域に少なくとも k 人のユーザが存在するように匿名化することで、個人を一意に特定させない手法である。位置情報を曖昧化することにより、位置情報を利用したサービスの品質低下が予想されるが、既存研究では、k 人以上のユーザが存在する領域の面積をできるだけ最小化することによって、品質低下を防いでいる。だが、提供するサービスの種類によっては、ユーザが存在する可能性がある位置の領域が最小化されることよりも、ユーザがどのような属性を持つ場所にいるか（たとえば、繁華街にいるのかビジネス街にいるのか）についての情報が失われないことのほうが重要である場合もある。本論文では、場所の属性に基づいた匿名化を行うことにより、場所の属性を利用するサービスの品質低下を防ぐ手法を提案する。また、場所の属性を設定する手法として、Web 情報を用いた手法を提案する。

1. はじめに

ユーザの位置情報を GPS 機能がついた携帯電話等から取得し、その情報に基づいてサービスを提供するロケーションベースサービスが普及しつつある。ユーザの位置情報を取得し、位置情報に応じて広告などの情報を通知するサービス [5] が考えられる。また、香港の企業が、ユーザの日々の位置情報を取得し、他のサービス提供者に向けて公開している例もある [7]。ユーザの正確な位置情報が悪意のある第三者に渡ると、ユーザの居住地や職場が特定されてしまう脅威が存在する。したがって、位置情報を適切に管理できるサービス提供者でなければ、ロケーションベースサービスを提供することはできない。だが、位置情報を適切に管理できるサービス提供者が、位置情報からユーザを一意に特定できないように匿名化を施すことによって、匿名化された位置情報を第三者に提供することが可能となる。匿名化された位置情報は、他の事業者が位置情報に基づいたサービス提供を行ったり、データマイニングを行う研究のために利用することができる。

ユーザを一意に特定できないように匿名化を行う指標の一つとして、k-匿名性が提案されている [9, 8]。たとえば、ユーザの位置が (x, y) であるとする。$x_1 < x < x_2, y_1 < y < y_2$ を満たす (x_1, x_2, y_1, y_2) を用意し、ユーザが (x_1, x_2, y_1, y_2) の頂点によって表される矩形領域に存在するという情報のみを第三者に提供することにより、ユーザの正確な位置を隠蔽する。このとき、x_1, x_2, y_1, y_2 の値は、(x_1, x_2, y_1, y_2) の領域に k 人以上のユーザが存在するように設定する。この領域内に k 人以上のユーザが存在するので、この情報を通知する元になったユーザを第三者は一意に特定することができない。k-匿名性を対象とする既存研究は、k 人以上のユーザが存在する領域の面積をできるだけ最小化することをめざしている。k 匿名性を満たす領域の面積を最小化する問題は、NP 困難であることが示されているため [6, 1]、より簡潔でありかつ良い解を与えるアルゴリズムが提案されている [3, 4]。

だが、提供するサービスの種類によっては、ユーザが存在する可能性がある位置の領域が最小化されることよりも、ユーザがどのような属性を持つ場所にいるか（たとえば、繁華街にいるのかビジネス街にいるのか）についての情報が失われないことのほうが重要である場合もある。たとえばある地域において、駅の東側と西側で様相が異なる街を考える。ユーザがどのような属性を持つ場所にいるかの情報を考慮せずに匿名化を行うと、ユーザが駅の近くにいる場合、ユーザが駅の東側にいるのか西側にいるのかの情報が失われる恐れがある。一方、ユーザが駅の東側にいるか西側にいるかの情報を追加情報として第三者に提供すると、k 匿名性が失われる恐れがある。したがって、ユーザがどのような属性を持つ場所にいるかの情報が第三者にとって重要である場合は、この属性情報を考慮した上で k 匿名性を満たすアルゴリズムを提案する必要がある。

本論文では、場所の属性に基づいた匿名化を行うことにより、場所の属性を利用するサービスの品質低下を防ぐ手法を提案する。また、場所の属性を設定する手法として、Web 情報を用いた手法を提案する。後者の提案については、マピオン（図 1）の Web サイトを利用した。マピオンの Web サイトには、各駅の周辺に、飲食店や公園等がそれぞれいくつ存在しているかの値が記載されている。このデータを用いて駅のクラスタリングを行うことにより、駅周辺の状況（ビジネス街、住宅街、繁華街等）の把握を行った。

2. 場所属性情報を考慮した匿名化

2.1 場所属性の記述方法の定義

場所の属性の集合を C、各属性を C_i とおく。たとえば、場所の属性としてビジネス街、住宅街、繁華街の 3 つが想定されている場合は、C = { ビジネス街、住宅街、繁華街 } であり、C_1 = ビジネス街 となる。地点 (x, y) が属性 C_i を保有している度合を $V_i(x, y)$ と表現する。地点 (x, y) がビジネス街

連絡先: 清雄一, (株) 三菱総合研究所, 東京都千代田区大手町 2-3-6, ysei@mri.co.jp

図 1: マピオン

に完全に属している場合は、$V_1(x,y)=1$ であり、全く属していない場合は $V_1(x,y)=0$ である。このように場所の属性情報として、ある地点 (x, y) がある属性を保持しているかどうか 2 値で表されている場合以外に、連続値で表現されている場合も考えられる。この場合は、$V_1(x,y)$ の値は 0 から 1 までの実数値を取る。

2.2 匿名化指標の提案

多くの既存研究はデータを曖昧化することによって k 匿名性を実現しており、曖昧化されたデータの有用性を失わないようにするため、曖昧化の幅をできるだけ小さくすることを目標としている [4, 2]。たとえば、あるユーザの位置が (x,y) であるとするとき、k 人以上が存在し、かつ面積が最小となるような矩形の領域を算出し、その頂点 (x_1,x_2,y_1,y_2) を第三者に提供する。一方、本論文においては、面積が最小となる領域を算出することが目標ではなく、場所属性も考慮した上で算出する必要があるため、新たな指標が必要である。

矩形の面積をできるだけ小さくするという目標に加え、矩形の頂点を (x_1,x_2,y_1,y_2) としたとき、その中に含まれる各点 (s,t) において、各場所属性の値 $V_i(s,t)$ の最大値と最小値の幅を小さくすることも目標となる。前者と後者のどちらを重要視するかは、曖昧化されたデータを利用する第三者が設定する。

したがって、(x_1,x_2,y_1,y_2) で表現される領域を S とおき、最小化すべき指標を T とすると、

$$T = |x_2 - x_1| + |y_2 - y_1| + w \cdot \sum_{i=0}^{|C|-1} \left\{ \max_{(s,t)\in S} V_i(s,t) - \min_{(s,t)\in S} V_i(s,t) \right\} \quad (1)$$

となる。S 内に k 人以上のユーザが存在し、かつ、式 1 を最小化するように、曖昧化された領域 S を設定する。

3. Web 情報を用いた場所属性の抽出

マピオンの Web サイトには、各駅の周辺に、飲食店や公園等がそれぞれいくつ存在しているかの値が記載されている。このデータを用いて駅のクラスタリングを行う。クラスタリングを行った結果に対して、それぞれの属性（ビジネス街、住宅街等）の付与については、手動で行う。

3.1 マピオン Web サイトからのデータ収集

Mapion 電話帳には、「和食店」や「スーパーマーケット」などの各カテゴリについて、駅ごとにその施設についての情報が

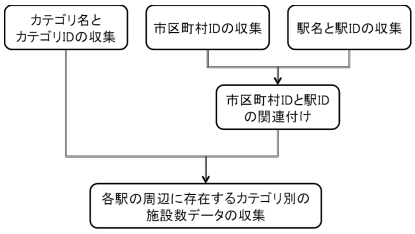

図 2: 駅単位のカテゴリ別施設データについての収集手順

記載されている。Mapion 電話帳から、この情報を収集し、駅ごとにおけるカテゴリ別施設数のデータを作成した。図 2 にデータ作成の手順を示す。(1) カテゴリ名とカテゴリ ID の収集、(2) 市区町村 ID の収集、(3) 駅名と駅 ID の収集では、マピオン Web サイトにおけるページ構造の把握を行っている。

1. カテゴリ名とカテゴリ ID の収集 Mapion 電話帳では、「和食店」などのカテゴリに分類して情報を記載している。このカテゴリ名の収集を行った。また、各カテゴリには ID が付加されており、この ID に基づいて、カテゴリ用の URL が決定されている。そのためカテゴリ ID の収集も同時に行った。たとえば、「和食店」のカテゴリ ID は「M01001」であり、「和食店」に関する情報は、`http://www.mapion.co.jp/phonebook/M01001/` に記載されている。

2. 市区町村 ID の収集 Mapion 電話帳では、「千代田区」など市区町村ごとに分類して情報を記載している。また、各市区町村には ID が付加されており、この ID に基づいて、市区町村用の URL が決定されている。そのため市区町村 ID の収集も同時に行った。たとえば、「千代田区」の市区町村 ID は「13101」であり、「千代田区」に関する「和食店」の情報は、`http://www.mapion.co.jp/phonebook/M01001/13101/` に記載されている。

3. 駅名と駅 ID の収集 Mapion 電話帳では、「大手町駅」などの駅ごとに分類して情報を記載している。それらの駅名の収集を行った。また、各駅には ID が付加されており、この ID に基づいて、駅用の URL が決定されている。そのため駅 ID の収集も同時に行った。たとえば、「大手町駅」の駅 ID は「ST22564」であり、「大手町駅」に関する「和食店」の情報は、`http://www.mapion.co.jp/phonebook/M01001/13101/ST22564/` に記載されている。ここで、「大手町駅」は「千代田区」に存在するため、URL には「千代田区」の市区町村 ID である「13101」も含まれている。

4. 市区町村 ID と駅 ID の関連付け上記 (3) で述べたように、各駅がどの市区町村に属するかの情報が必要となる。従って、各駅がどの市区町村に存在するかの関連付けを行った。

5. 各駅の周辺に存在する各カテゴリの施設数データ収集上記の (1) (2) (3) をまとめると、市区町村 A_i に属する駅 S_j の周辺に存在するカテゴリ C_k の施設に関する情報は、`http://www.mapion.co.jp/phonebook/`C_k`/`A_i`/`S_j`/` から得られる。ただし、この URL からは、駅 S_j の周辺に存在するカテゴリ C_k の施設についての情報を、最大 20 件までしか

表 1: クラスタリングに使用した施設カテゴリ

カテゴリ番号	カテゴリ	説明
1	グルメ	グルメ施設数合計
2	デパート・百貨店	デパート数
3	スーパー・コンビニ	スーパーの数とコンビニの数の合計
4	本屋・文房具	本屋・文房具店数
5	電気屋・家電	電気屋・家電店数
6	カラオケ・ゲームセンター	カラオケ・ゲームセンター数
7	美術館・博物館	美術館・博物館数
8	公園・緑地	公園・緑地数
9	ホテル	ホテル数
10	ラブホテル・モーテル	ラブホテル・モーテル数
11	美容院・マッサージ	美容院・マッサージ店数
12	病院	病院数
13	大学・短大	大学・短大数
14	学校	小学校・中学校・高等学校数
15	官公庁	官公庁施設数
16	大使館・領事館	大使館・領事館数
17	ビジネス（写真・デザイン・装飾・録音）	写真・デザイン・装飾・録音関連の事業者数
18	ビジネス（建設・工事業）	建設・工事業関連の事業者数
19	ビジネス（港湾・海運）	港湾・海運関連の事業者数
20	ビジネス（倉庫業）	倉庫業関連の事業者数
21	ビジネス（水産・畜産業）	水産・畜産業関連の事業者数
22	ビジネス（アパレル業服飾雑貨・趣味・民芸・工芸品）	アパレル業服飾雑貨・趣味・民芸・工芸品関連の事業者数

得られない。21 件目以上の情報を取得したい場合は、この URL の最後に”[num].html”を付加する。ここで、[num] には、1 以上の整数が入り、$20 \times ([num]-1)+1$ 件目から $20 \times [num]$ 件目までの施設情報を取得することができる。たとえば $[num]=2$ の場合は 21 件目から 40 件目までの情報を取得することができ、$[num]=3$ の場合は 41 件目から 60 件目までの情報を取得することができる。従って、市区町村 A_i に属する駅 S_j の周辺に存在するカテゴリ C_k の施設数を取得する場合は、まず `http://www.mapion.co.jp/phonebook/`C_k`/`A_i`/`S_j`/` のページの情報を取得し、このページに何件の施設情報が記載されているかを確認する。20 件未満である場合は、その件数が駅 S_j の周辺に存在するカテゴリ C_k の施設数となる。20 件の施設情報が記載されている場合は、次に `http://www.mapion.co.jp/phonebook/`C_k`/`A_i`/`S_j`/2.html` のページ情報を取得し、このページに何件の施設情報が記載されているかを確認する。20 件未満である場合は、その値に 20 を足した値が駅 S_j の周辺に存在するカテゴリ C_k の施設数となる。20 件の施設情報が記載されている場合は、さらに `http://www.mapion.co.jp/phonebook/`C_k`/`A_i`/`S_j`/3.html` のページ情報を取得する。以降これを繰り返す。この処理を、収集したカテゴリ ID の集合と駅 ID の集合の組み合わせに対して順次行い、各駅の周辺に存在する各カテゴリの施設数についてのデータを作成した。

3.2 駅のクラスタリング

Mapion 電話帳から取得した駅単位のカテゴリ別施設数のデータを用いて、駅単位の街のクラスタリングを行った。用いたデータは Mapion 電話帳に掲載されている施設数のデータを、表 1 に挙げた項目について集計・集約を行ったものである。分析対象となった駅は首都圏の 1547 の駅である。

クラスタリングの前処理としてデータの正規化を行った。これはカテゴリ毎の全体の施設数の総数の偏り（ex. 大使館数はコンビニエンスストア数と比較してそもそもの施設数が少ない、等）を考慮するためである。正規化の方法は以下である。

$$\text{正規化後施設数} = \frac{\text{正規化前施設数} - \text{カテゴリ内最小施設数}}{\text{カテゴリ内最大施設数}} \tag{2}$$

正規化後のデータを k-means アルゴリズムによりクラスタリングを行った。クラスタ数は 25 とした。

3.3 クラスタリング結果と考察

Mapion 電話帳から取得した駅単位のカテゴリ別施設数のデータを用いて、街のクラスタリングを行った結果を以下に示す。

表 2: 特徴的なクラスタ

駅	クラスタ番号
霞ヶ関駅（東京都）	13
国会議事堂前駅	13
永田町駅	13
溜池山王駅	13
虎ノ門駅	13

駅	クラスタ番号
霞ヶ関駅（埼玉県）	18
西川越駅	18
的場駅	18
花崎駅	18
元加治駅	18
新座駅	18
西大家駅	18
川角駅	18
東毛呂駅	18
毛呂駅	18
作草部駅	18
天台駅	18
船橋日大前駅	18
京成大久保駅	18
物井駅	18
千葉ニュータウン中央駅	18
市ヶ谷駅	18
玉川学園前駅	18

駅	クラスタ番号
御茶ノ水駅	18
三田駅（東京都）	18
早稲田駅	18
西早稲田駅	18
水道橋駅	18
茗荷谷駅	18
東大前駅	18
駒沢大学駅	18
三軒茶屋駅	18
成城学園前駅	18
中野坂上駅	18
駒込駅	18
江古田駅	18
新桜台駅	18
片倉駅	18
八王子みなみ野駅	18
大塚・帝京大学駅	18
中央大学・明星大学駅	18

駅	クラスタ番号
南町田駅	18
新小金井駅	18
鷹の台駅	18
多摩動物公園駅	18
程久保駅	18
新秋津駅	18
秋津駅	18
稲城駅	18
花月園前駅	18
国道駅	18
新高島駅	18
日吉駅（神奈川県）	18
新百合ヶ丘駅	18
六会日大前駅	18
箱根板橋駅	18
淵野辺駅	18
東海大学前駅	18

駅	クラスタ番号
大久保駅（東京都）	21
新大久保駅	21
東新宿駅	21
鶯谷駅	21
京成上野駅	21
浅草[つくばEXP]駅	21
蒲田駅	21
神泉駅	21
大塚駅（東京都）	21
大塚駅前駅	21
伊勢佐木長者町駅	21
阪東橋駅	21
黄金町駅	21

表 3: クラスタごとのカテゴリ別施設数の平均

クラスタ	街カテゴリ	グルメ	デパート	スーパー・コンビニ	本屋・文房具	電気屋・家電	カラオケ・ゲームセンター	美術館・博物館	公園・緑地	ホテル	ラブホテル・モーテル	美容院・マッサージ
1	ビジネス繁華街	1193.78	0.87	80.11	35.87	43.33	10.87	8.58	0.89	38.87	1.44	129.87
2	住宅街	99.37	0.03	11.88	3.78	4.84	1.88	3.50	1.37	12.13	0.18	28.58
3	住宅街	82.79	0.01	11.93	4.09	3.83	1.00	0.27	0.97	0.78	0.11	31.10
4	小規模繁華街	314.13	0.53	109.47	10.87	11.87	8.53	1.87	1.33	7.53	0.53	108.87
5	大規模住宅街	318.15	0.29	80.93	15.00	20.39	8.24	3.49	5.24	4.80	0.54	154.44
6	住宅街	44.44	0.00	9.45	2.05	3.85	0.41	2.88	0.98	1.38	0.08	22.15
7	住宅街	189.03	0.10	24.78	9.87	14.53	3.80	1.12	2.42	3.31	0.35	92.74
8	ビジネス街	495.95	0.30	35.10	22.10	27.43	5.25	1.83	0.55	15.85	0.30	81.58
9	住宅街	140.50	0.03	30.18	7.90	13.55	2.89	2.12	8.90	2.21	0.28	77.57
10	中規模繁華街	581.98	1.94	88.19	21.10	30.42	12.13	3.19	3.94	18.83	1.58	220.44
11	住宅街	78.15	0.02	15.18	4.23	8.93	1.43	0.70	1.58	1.27	0.11	40.99
12	大規模繁華街	1507.00	7.80	134.80	31.27	39.80	28.73	2.87	0.47	37.87	3.80	238.40
13	官公庁・大使館街	440.20	0.00	24.40	8.80	13.20	2.40	0.40	0.00	4.20	0.00	29.20
14	官公庁・大使館街	235.10	0.00	24.10	8.80	12.70	1.70	1.80	0.50	10.20	0.00	89.00
15	住宅街	18.85	0.00	3.89	0.71	1.88	0.28	0.25	0.53	0.81	0.08	7.98
16	ビジネス街	873.87	0.73	41.80	21.13	94.40	11.93	1.13	0.00	27.87	0.13	79.33
17	倉庫街	78.03	0.00	14.33	3.07	8.80	0.90	0.50	1.40	3.90	0.17	25.57
18	学生街	128.21	0.15	22.45	7.94	10.23	2.17	1.75	2.57	2.28	0.17	52.79
19	中規模繁華街	734.30	3.48	102.48	29.35	38.28	12.83	8.30	3.52	20.43	1.09	290.85
20	住宅街	149.20	0.87	43.93	9.40	12.27	4.27	4.73	19.07	1.07	0.07	77.00
21	歓楽街	388.38	0.38	42.00	12.89	25.00	8.23	1.85	0.54	41.85	8.31	110.54
22	小規模繁華街	294.32	0.39	39.94	14.28	19.78	8.59	1.48	2.82	8.42	0.58	144.98
23	倉庫街	195.75	0.00	30.17	7.92	23.75	4.25	2.42	2.17	5.75	0.42	57.93
24	中規模住宅街	198.81	0.10	22.40	8.54	9.82	3.81	0.73	0.73	4.44	0.38	71.70
25	中規模住宅街	239.38	0.28	30.59	10.77	18.10	2.79	7.38	4.00	8.87	0.58	102.92

クラスタ	病院	大学・短大	学校	官公庁	大使館・領事館	写真・デザイン・装飾・録音	建設・工事業	港湾・海運	倉庫業	水産・畜産業	アパレル業服飾雑貨・趣味・民芸・工
1	127.87	0.44	1.78	117.44	0.11	20.78	88.11	214.87	9.87	98.58	23.33
2	19.11	0.21	3.34	4.87	0.05	2.13	7.97	28.34	0.47	1.28	0.55
3	24.28	0.08	2.54	1.18	0.02	3.54	5.88	27.94	0.05	0.31	0.52
4	70.33	0.47	2.87	2.20	0.13	5.87	29.87	53.73	1.20	0.53	1.07
5	118.34	1.48	15.95	3.05	0.00	14.83	21.78	118.49	0.37	0.37	2.98
6	13.94	0.13	2.42	0.79	0.04	2.87	4.28	23.34	0.13	0.99	0.81
7	48.51	0.07	5.99	1.81	0.00	8.89	10.31	73.78	0.22	0.95	1.41
8	73.88	1.15	2.23	12.80	0.28	10.78	79.18	102.55	4.00	14.03	3.10
9	52.12	0.43	10.19	1.53	0.00	11.78	12.05	81.80	0.29	0.31	2.74
10	121.58	0.58	9.12	8.19	0.00	18.80	24.31	139.31	0.38	1.80	3.89
11	27.48	0.23	5.80	2.28	0.00	4.75	8.81	41.99	0.22	0.22	0.93
12	149.20	0.40	0.80	7.53	0.33	14.33	92.00	99.53	5.13	10.40	3.20
13	80.80	0.00	0.80	704.80	1.80	4.40	34.00	47.40	8.20	52.80	2.80
14	50.00	0.80	3.80	2.80	4.70	3.30	88.70	81.10	4.20	9.50	2.80
15	4.98	0.02	1.18	2.51	0.00	1.83	1.38	11.09	0.37	0.83	0.48
16	80.40	0.20	2.27	4.80	0.07	11.80	57.53	131.80	4.53	13.40	5.33
17	14.83	0.07	2.87	5.10	0.07	8.57	8.70	33.57	19.20	5.10	9.80
18	39.02	2.74	8.34	1.28	0.04	8.32	13.51	48.92	0.47	0.77	1.02
19	178.87	0.87	15.09	9.87	0.00	22.43	41.13	187.93	0.00	0.28	3.74
20	57.47	0.80	12.47	0.87	0.00	10.80	11.73	81.07	0.00	0.40	2.07
21	88.15	0.00	2.85	4.77	0.23	12.38	38.48	105.82	0.38	3.31	2.00
22	79.48	0.23	8.70	2.48	0.00	13.71	13.90	101.55	0.14	0.35	2.88
23	47.42	0.33	8.33	1.92	0.33	12.83	23.33	101.33	3.75	11.92	21.00
24	48.50	0.08	2.12	1.12	0.02	4.85	14.32	41.31	0.15	0.81	0.58
25	85.72	0.28	7.87	8.85	0.05	10.82	18.44	78.92	0.51	2.74	1.79

特徴的なクラスタを表 2 に、クラスタごとのカテゴリ別平均施設数を表 3 に示す。クラスタ 13 は、霞ヶ関（東京都）、国会議事堂前、永田町、溜池山王、虎ノ門が属しており官公庁施設数が他のクラスタと比較して高い。一方、クラスタ 18 は御茶ノ水、三田、日吉等が属しており大学・短大数が他のクラスタと比較して高い。また、クラスタ 21 は新大久保、神泉、蒲田等が属しており、宿泊施設の数が他のクラスタより多い。

このように同一クラスタ内には類似した街が分類される。クラスタごとのカテゴリ別平均施設数のデータよりクラスタに特徴付けを行った。街ごとに街の特徴を示すカテゴリ「街カテゴリ」を付与した。前述の特徴的なクラスタの例では、クラスタ 13 は官公庁街、クラスタ 18 は学生街、クラスタ 21 は歓楽街といった街カテゴリを付与した。表 4 に、付与した街カテゴリの一覧とクラスタごとの街カテゴリを示す。

3 つのクラスタを地図上にプロットしたものが図 3 である（地図は Google Earth を使用）。青色アイコンが官公庁街、赤色アイコンが学生街、黄色アイコンが歓楽街を示す。都心に官公庁街が位置し、山手線の西部〜北部に歓楽街が位置し、学生街が都心〜郊外に位置しているのが分かる。

3.4 クラスタリングを行わない場所属性設定方法

本章では、ある地域に存在する施設情報から、地域のクラスタリングを行った。クラスタリングした結果として、官公庁街・学生街等の属性を付与したが、実際には、第三者が有用であると想定するような属性の付与を行う。また、第三者が、曖昧化されたデータを受け取る時点では、どのような属性を付与すべきか定まっていない状況も考えられる。このような場合には、より汎用的な属性付与方法を利用することが考えられる。

表 4: 街カテゴリ一覧とクラスタごとの街カテゴリ

街カテゴリ
ビジネス街
ビジネス繁華街
大規模繁華街
中規模繁華街
小規模繁華街
大規模住宅街
中規模住宅街
住宅街
歓楽街
倉庫街
官公庁・大使館街

クラスタ	街カテゴリ
1	ビジネス繁華街
2	住宅街
3	住宅街
4	小規模繁華街
5	大規模住宅街
6	住宅街
7	住宅街
8	ビジネス街
9	住宅街
10	中規模繁華街
11	住宅街
12	大規模繁華街
13	官公庁・大使館街
14	官公庁・大使館街
15	住宅街
16	ビジネス街
17	倉庫街
18	学生街
19	中規模繁華街
20	住宅街
21	歓楽街
22	小規模繁華街
23	倉庫街
24	中規模住宅街
25	中規模住宅街

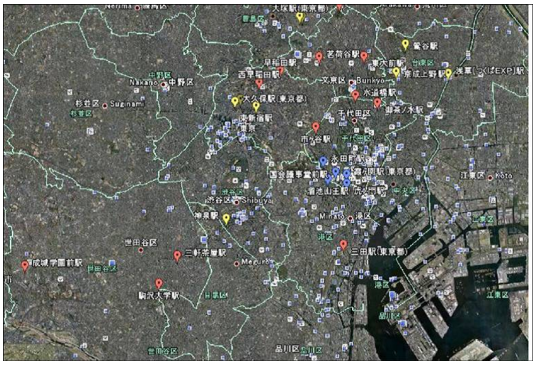

図 3: 街クラスタのプロット（官公庁街・学生街・歓楽街）

たとえば、グルメ施設やカラオケ施設等の施設それぞれについて、それらが存在する密度情報を属性として付与することができる。つまり、ある地域におけるグルメ施設の存在密度を属性 C_1 などとする。これにより、あるユーザがグルメ施設が多い地域にいるのか、カラオケ施設が多い地域にいるのかの情報をできるだけ失わないように匿名化を行うことが可能となる。

各施設が存在する密度情報の算出方法として、各施設が存在する地点を中心に二次元正規分布を描き、その和を取ることが考えられる。ある施設が地点 (μ_x, μ_y) に存在する場合、地点 (x, y) へ与える属性値の大きさは

$$N(x, y, \mu_x, \mu_y) = \frac{1}{2\pi} \exp\left(-\frac{(x-\mu_x)^2(y-\mu_y)^2}{2}\right) \quad (3)$$

となる。属性 C_i を与える施設が n 個存在し、それぞれの位置が (μ_{x_j}, μ_{y_j}) である場合、地点 (x, y) における属性値 $V_i(x, y)$ は、

$$V_i(x, y) = \sum_{j=0}^{n-1} N(x, y, \mu_{x_i}, \mu_{y_i}) \quad (4)$$

となる。

4. おわりに

本論文では、場所の属性情報を用いた k 匿名化手法の提案を行った。また、場所の属性情報の決定手法として、Web 情報を用いた手法を提案し、実際にマピオン Web サイト上の情報を用いて、駅のクラスタリングを行った。今後、マピオン Web サイト等から得られた情報を用いて場所の属性情報を細かく設定し、提案した k 匿名化手法についての評価を行う予定である。

参考文献

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. *Database Theory-ICDT 2005*, pp. 246–258.

[2] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE '05: Proceedings of the 21st International Conference on Data Engineering*, pp. 217–228, Washington, DC, USA, 2005. IEEE Computer Society.

[3] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, p. 60. ACM, 2005.

[4] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pp. 25–25, 2006.

[5] J. Meyerowitz and R. Roy Choudhury. Hiding stars with fireworks: location privacy through camouflage. In *MobiCom '09: Proceedings of the 15th annual international conference on Mobile computing and networking*, pp. 345–356, New York, NY, USA, 2009. ACM.

[6] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, p. 228. ACM, 2004.

[7] M. Terrovitis and N. Mamoulis. Privacy preservation in the publication of trajectories. In *IEEE International Conference on Mobile Data Management*, pp. 65–72, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

[8] R. Yarovoy, F. Bonchi, L. Lakshmanan, and W. Wang. Anonymizing moving objects: how to hide a MOB in a crowd? In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 72–83. ACM, 2009.

[9] P. Zacharouli, A. Gkoulalas-Divanis, and V. S. Verykios. A k-anonymity model for spatio-temporal data. In *ICDEW '07: Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, pp. 555–564, Washington, DC, USA, 2007. IEEE Computer Society.