自由記述データを用いたテキストマイニングによる都市のイメージ分析の有用性

群馬工業高等専門学校　環境都市工学科　学生会員　○入澤　覚
群馬工業高等専門学校　環境都市工学科　正会員　森田　哲夫
群馬工業高等専門学校　専攻科　生産システム工学専攻　学生会員　藤橋　貞光
群馬工業高等専門学校　電子情報工学科　荒川　達也

1. はじめに

(1) 研究の背景・目的

都市計画分野において、住民などの都市のイメージを把握することは重要である。都市のイメージを把握する方法としてはアンケート調査が一般的であるが、従来の SD 法などでは設問が限られているため、多様な意見を抽出することが難しい。またアンケート調査では併せて自由記述欄を設け、自由意見を記入してもらうことが多いが、これまでは分析観点の恣意性から十分な検討がされてこなかった。近年、近年言語処理技術の発達により大規模なテキストデータの定量的な分析が可能になった。しかし、テキストマイニング手法による都市イメージの分析方法は、まだその有用性は明らかになっていない。

本研究は、アンケート調査の自由記述データを用いたテキストマイニングによる都市のイメージ分析の有用性を検証することが目的である。

(2) 既存研究と本研究の位置づけ

都市のイメージを分析する研究としては数多くあるが、そのなかでも本研究に関連するものは、角野[1)]、斉藤ら[2)]の研究がある。自由記述データに対してテキストマイニングを用いて分析している研究としては、大塚ら[3)]が水・緑環境に着目して前橋市の都市のイメージ分析を行った。また長塩ら[4)]は個人属性・地区特性と都市イメージの関係性を分析した。また永野[5)]らは自由記述データとプリコードデータを比較分析して両者の関係性を把握し、自由記述データの分析方法を検討した。佐々木ら[6)]はブログマイニングから行動データを抽出し、アンケート調査との比較を行った。さらに佐々木ら[7)]はテキストマイニング手法を用いて、まちづくりのワークショップの討議内容の視覚化、把握のための研究を行った。本研究では自由記述データを定量的に分析することで、都市イメージの構造を定量的・視覚的に明らかにし、有用性を検証する。

(3) 分析仮説

テキストマイニングを用いた自由記述データからの都市イメージ分析の有用性を示すため、次の 3 つを分析仮説とし、それぞれ検証するための分析を行う。1)自由記述データから抽出された単語が、プリコードデータと関連している。2)プリコードデータでは、設問に対する評価値が得られるが、テキストデータから都市イメージ分析を行うことで満足の質が分析できる。3)自由記述データから、プリコードの評価項目には想定していなかったキーワードを把握できる。

2. 分析方法

(1) 対象地域の設定

対象地域は群馬県前橋市の利根川左岸の地域とした。前橋市は「水の郷 100 選」の一つに選ばれており、市も「水と緑と詩の町」を標榜としている。対象地域には利根川、広瀬川、桃の木川を始めとする河川や大正用水などの用水路が数多く存在することや、前橋公園や敷島公園などの大規模公園が存在するため、水、緑環境が都市のイメージに与える影響があると考えた。

(2) アンケート調査の概要

アンケート調査の企画概要を表 1 に示した。対象地域内の 4,000 世帯（抽出率 9.5%）を対象に調査を実施し、2,118 票を回収した。本分析では自由記述欄に記入のある 1,614 票をそれぞれ有効票とした。また自由記述を促す指示文で、「前橋の良い点、好きなところ」と限定したのは、「良い点」と指定せずに「前橋市のイメージ」を問うことにより、居住者すなわち当事者としてはネガティブなイメージに焦点が当たりやすくなることを想定したためである。

(3) 使用する分析手法

自由記述データなどのテキストデータは、何らかの方法で客観的・定量的に解析できるように加工する必要がある。自然言語処理分野では、いくつかの解析手

キーワード　イメージ　テキストマイニング　コレスポンデンス分析
連絡先〒371-8530 前橋市鳥羽町 580 群馬工業高等専門学校環境都市工学科 TEL027-254-9179E-mail:tmorita@civil.gunma-ct.ac.jp

表1　アンケート企画概要

調査期間	配布：2008年11月上旬～下旬 配布：2008年12月～2009年2月
調査方法	配布：調査員によるポスティング 回収：郵送回収
配布世帯数	対象地域4,000世帯（抽出率9.5%）の構成員2票ずつ
調査内容	1)個人属性（性別年齢、普段利用する交通手段、職業、家族人数、住宅種類） 2)生活の質の主観的評価値（20項目＋総合評価、5段階評価：1.満足、2.やや満足、3.どちらでもない、4.やや不満、5.不満.） 3)自由記述「前橋の良い点、好きなところ、場所などを短い言葉や文で自由に書いてください」 世帯：1,293世帯(32.3%)
回収数	個人：2,118票回収、うち自由記述欄に記入あり1,614票

法やそのためのソフトウェアが開発されており、本研究は、形態素解析には技術情報が公開されている KH Coder[8]を使用することとした。また分析手法はコレスポンデンス分析、共起ネットワーク分析を用いる。これらはそれぞれ4、5章で説明する。

(4)コーディングルールの作成

自由記述データは、すべてのサンプルが同じ表現方法で文章を記述していることは考えられない。よって同じ意味の表現などを一つのコードとしてまとめる必要がある。今回は以下の点に従ってコーディングルールを作成した。

a.「みどり」や「緑」など、同じ意味で表記の仕方が異なるものに同一のコードを与えた。

b.「多い」や「たくさん」など同じ意味でも表現が異なるものに同一のコードを与えた。

c.固有名詞について、「覚満渕」、「臨江閣」などは「覚」、「満」、「臨」などというように単語が分解されたため、それらには固有名詞のコードを与えた。

コーディングルールの作成に再現性を持たせるため、類義語は類義辞典（シソーラス）[9]を参考にしてまとめた。また、調査票から実際の単語の使われ方を確認し、それぞれ単語の意味を決定してルールを作成した。

3. 語の出現頻度の基礎集計

アンケート調査で得られた自由記述データに形態素解析を行った。最頻出50語とその出現数を表2に示す。

表2　自由記述データ最頻出50語

抽出語	出現数	抽出語	出現数	抽出語	出現数
多い	408	思う	142	医療	69
良い	402	自然	137	便利	68
公園	389	好き	127	環境	67
緑	320	広瀬川	116	川	66
無い	290	利根川	89	薔薇	63
する	289	場所	88	赤城山	62
ある	277	比較的	84	なる	62
敷島	229	県庁	83	周辺	60
少ない	217	落着き	83	生活	58
近辺	214	見る	83	地域	58
水	181	とても	82	景色	58
前橋	171	おいしい	80	都市	53
災害	168	道路	80	車	52
住む	162	山	79	空気	51
街	149	恵まれる	77	特に	50
易い	146	きれい	73	楽しい	50
静か	144	出来る	70	以上50語	

4. プリコードデータと自由記述データの関係の分析

(1)分析方法

一つ目の仮説を検証するための分析を行う。アンケートの設問の5段階評価値と、テキストデータから抽出した単語の出現数の関係をクロス集計したものにコレスポンデンス分析を行い、設問の5段階評価値と自由記述データの抽出語の対応関係を示す。この分析では関連の高い語同士は近くにプロットされるが、この距離は相対的なものである。そのためこの分析のみでは距離が遠い、近いという判別に客観性がない。そこでコレスポンデンス分析の結果にクラスター分析を行い、同じクラスターに所属するならば距離が近く、単語同士の関連性があるとした。

(2)分析結果

この分析で使用した単語は、自由記述データの最頻出50語である。アンケートの設問項目「身近な緑に恵まれている」についての分析結果を図1に示す。図1中の緑色の枠は11のクラスターに分類した場合であり、青色の枠は5つのクラスターに分類した場合である。また、この設問の5段階評価値の平均値は2.083（数値の小さい方が高い評価）であり、やや満足しているという評価であった。

11のクラスターに分類した結果では、「満足」付近には「空気」「とても」がプロットされた。「やや満足」付近には「住む」「生活」「水」「薔薇」「おいしい」「比

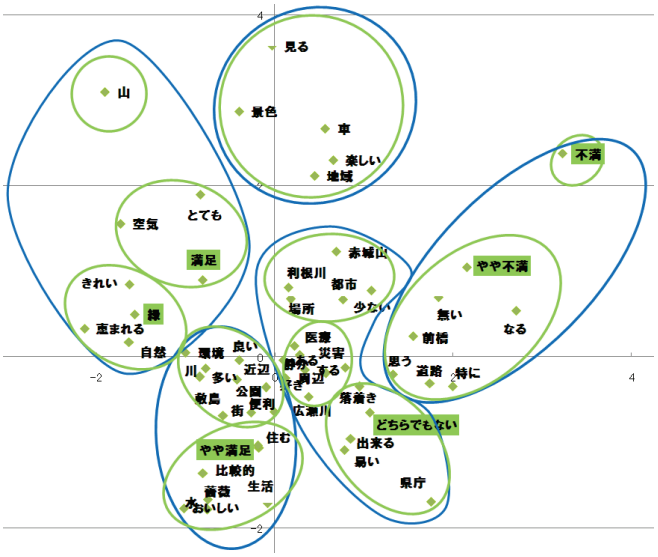

図１　コレスポンデンス分析結果（身近な緑に恵まれている）

較的」がプロットされた。「どちらでもない」付近には「落着き」「易い」「出来る」「県庁」がプロットされた。「やや不満」近くには「無い」「なる」「前橋」「思う」「道路」「特に」がプロットされた。「不満」付近には何もプロットされなかった。

「満足」～「不満」のいずれにも、「身近な緑に恵まれている」という設問に合致する単語はプロットされなかった。11 のクラスターに分類した結果では、設問の 5 段階評価値と自由記述データの抽出語の明確な対応関係を示すことができなかった。

5 つのクラスターに分類した結果では、「満足」が含まれる満足クラスター、「やや満足」が含まれるやや満足クラスター、どちらでもないクラスター、「やや不満」「不満」が含まれる不満クラスター、その他のクラスターの 5 つのクラスターに分類することができた。デンドログラムをより長い距離で切ることで、プロットされたより多くの単語が、満足から不満のどの評価に関連があるかどうかを分析する。

満足クラスターには「山」「空気」「とても」「きれい」「緑」「恵まれる」「自然」が所属している。やや満足クラスターには「住む」「生活」「水」「薔薇」「おいしい」「比較的」「良い」「環境」「近辺」「多い」「川」「公園」「便利」「町」「敷島」が所属している。どちらでもないクラスターには「出来る」「県庁」「易い」「落ち着き」「広瀬川」「周辺」「する」「周辺」「ある」「好き」「災害」「医療」「場所」「都市」「利根川」「赤城山」「少ない」「静か」が所属している。不満クラスターには「無い」「なる」「前橋」「思う」「道路」「特に」が所属している。

5 つのクラスターに分類した結果では、自然環境、生活環境に関するキーワードが満足、やや満足と関連していることが明らかになった。

5. 評価項目に対する満足の質の分析

(1) 分析方法

二つ目の仮説を検証するための分析を行う。プリコードデータのみの分析では設問に対する評価値を示すことはできるが、どのように良いのかという点についてはわからない。この分析では、テキストデータ用いることで、5 段階評価値の満足の質が分析可能か検討する。共起ネットワークにより単語間のつながりやパターンを明らかにする。次に、コレスポンデンス分析との結果と比較し、「満足」～「不満」付近にプロットされた単語がどのように形容されているかを分析する。

(2) 語の共起関係の算出

アンケート調査で得られた自由記述データから共起ネットワークを作成し、各単語の共起関係を算出する。共起関係の大きさには$Jaccard$係数を用いた。$Jaccard$係数とは、語ωと語ω'の共起関係を示し、大きいほどつながりが強いことを示す。$p(\omega \cap \omega')$は語ωと語ω'の共起確率を表し、$p(\omega \cup \omega')$は語ωと語ω'のいずれかが出現する確率を表す。語ωと語ω'の$Jaccard$係数は式(1)で与えられる。

$$Jaccard(\omega, \omega') = \frac{p(\omega \cap \omega')}{p(\omega \cup \omega')} \qquad (1)$$

(3) 共起ネットワーク分析

アンケートの自由記述データから、単語の共起関係を把握するため、つながりの強い語を結び、視覚的に表現する共起ネットワーク分析を行う。単語は頻出 200 語を分析対象とし、$Jaccard$係数 0.15 以上の共起ネットワークを示したものが図 2 である。この分析により、次のような語の共起関係が強いことがわかった。

「敷島公園」、「緑が多い」、「災害（が）少ない」、「物価（が）安い」、「住む（住み）易い」、「水（が）おいしい」、「風（が）強い」、「台風・地震・水害」、商業施設である「けやき（ケヤキ）ウォーク」、「けやき（の）街路樹」、「広瀬川畔」、「上毛三山」、「公共施設（が）充実」、「桃の木川（の）サイクリングロード」などである。

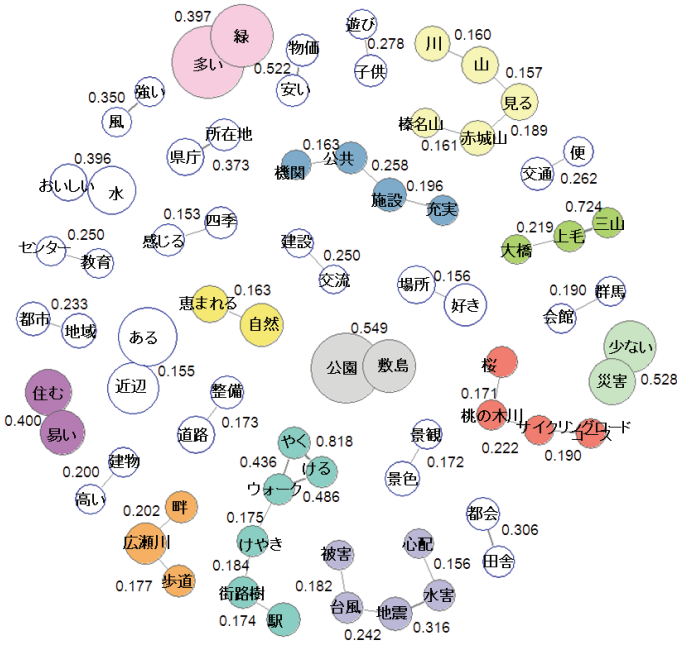

図 2　共起ネットワーク（Jaccard係数 1.5 以上）

表 3　名詞の形容のしかた（Jaccard係数 1.0 以上）

名詞	形容詞	Jaccard係数
災害	少ない	0.528
物価	安い	0.522
緑	多い	0.397
水	おいしい	0.396
風	強い	0.350
建物	高い	0.200
空気	きれい	0.148
便	悪い	0.143
地震	少ない	0.132
交通	悪い	0.131
道路	広い	0.117
山	美しい	0.109
四季	楽しい	0.107
野菜	安い	0.106
生活	易い	0.103

(4) 名詞の形容のしかた

名詞と形容詞のみで共起関係を算出し、名詞がどのように形容されているかを表 3 に示した。ここでは、4 章のコレスポンデンス分析の満足、やや満足クラスターに所属する単語についてみる。コレスポンデンス分析で満足という評価には「緑」「空気」「山」が関係していた。表 3 をみると、その満足の質は「緑（が）多い」、「空気（が）きれい」、「山（が）美しい」と形容されている。やや満足という評価には「生活」「水」が関係しており、その満足の質は「生活（が）し易い」、「水（が）おいしい」形容されている。

6. まとめ

(1) 本研究のまとめ

4 章のコレスポンデンス分析では「満足」、「やや満足」付近に自然環境に関する単語が出現しており、身近な緑に恵まれていることの満足度と評価が一致していると考えられる。すなわち、プリコードデータと自由記述データには関係性があり、自由記述データはプリコードデータで得られる結果の一部分を示していると考えられる。5 章では共起ネットワーク分析を用い、プリコード設問の満足さ、やや満足さに対する満足の質を示すことができたと考えられる。

(2) 今後の課題

本研究では自由記述データをプリコードデータと比較し、プリコードデータによる分析の結果の解釈の補助として、その有用性を見いだせた。今後は、自由記述データからより多くの定量的な情報を得られるかどうかを検討することが今後の課題である。

【参考文献】

1)角野幸博：地域メージの構成要素に関する研究-大阪府南北地域を事例に-，日本都市計画学会都市計画論文集，No.16，pp373-378，1981
2)斉藤和夫・石崎裕幸・田村亨・桝屋有三：都市のイメージ構造と地域特性の関係に関する研究，都市計画学会研究・論文集 No.14，pp467-474，1997
3)大塚裕子・森田哲夫・吉田朗・小島浩・塚田伸也：テキストマイニングによる都市・景観イメージ分析-水・緑環境に着目して-，土木計画学研究・講演集，Vol.41，CD-ROM（132），2010
4)長塩彩夏・森田哲夫・大塚裕子・塚田伸也：個人属性・地区特性と都市のイメージの関係に関する分析－水環境に着目して－，第 37 回土木学会関東支部技術研究発表会，講演概要集，CD-ROM，2010
5)永野峻祐・小根山裕之・大口敬・鹿田成則：形態素解析を用いたアンケート調査自由記述欄の分析手法に関する研究～路面電車利用意識調査データを用いたケーススタディ～，土木計画学研究・講演集，Vol.43，CD-ROM，2011
6)佐々木邦明・紀藤舞華・山崎慧太：ブログマイニングからの行動データの抽出・分析可能性とアンケート調査との比較，土木計画学研究・講演集，Vol.43，CD-ROM，2011
7)佐々木邦明・丸石浩一：テキストマイニングを用いたワークショップの討議内容の特徴把握と可視化に関する研究，都市計画論文集，Vol.46，No.3，2011
8)KH coder ウェブサイト，http://khc.sourceforge.net/，2012.1.16 閲覧
9)類語辞典・シソーラス - Weblio 辞書，http://thesaurus.weblio.jp/，2012.1.16 閲覧