地域に存在するテキストデータによる地域イメージの構成要素に関する分析

群馬工業高等専門学校　環境都市工学科　学会員　○牛口　聖矢
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　環境都市工学科　長塩　彩夏
前橋市都市計画部まちづくり課　正会員　塚田　伸也
（財）計量計画研究所　言語・行動研究室　大塚　裕子

１．はじめに

(1) 研究の背景・目的

地域の歴史や文化、イメージは、地域に存在する文章や文字に刻み込まれていると考えられる。都市・地域計画やまちづくりを検討する際、地域の歴史や文化による地域のまとまりを考慮することが重要である。また、近年、地域や都市のイメージ分析に、自然言語処理分野のテキストマイニング技術を適用する例がみられるようになっている。

本研究においては、地域の歴史や文化を刻み込んでいるテキストデータとして、学校の校歌の歌詞（テキストデータ）に着目し、以下の 2 つを目的に研究を進める。

a. 地域で共有するイメージを形成する語（キーワード）の抽出と、その語による地域のまとまりの把握
b. キーワードとその他の語との関係を分析することによる地域イメージの構成要素の把握

(2) 既存研究と本研究の位置づけ

地域や都市のイメージについては、70 年代以降いくつかの視点で数多くの研究がなされている。それらは、市民・住民が都市あるいは地区を評価し、多変量解析等により評価構造を明らかにしている研究が多い。斎藤ら[1]が、形容詞対による評価値と地域特性との関連性を分析している。これら研究は調査票に予め設定された選択肢から得られるデータを使用した分析である。一方、土木計画学分野、都市計画学分野では、参加型計画において得られる意見の分類・分析に自然言語処理技術を用いる研究[2)3)]、アンケート調査の自由記述データにテキストマイニングを適用し、地域特性との関係を分析した研究[4]が存在するが、処理結果の分析法等が確立されておらず、萌芽的な研究課題である。

本研究は、既存研究[4]の方法を継承し、校歌の歌詞を対象に、地域で共有するイメージを分析するものであり、萌芽的な研究課題であるテキストマイニングを適用し、地域のまとまりやその構成要素を把握する点が特徴である。また、都市・地域の計画課題との関係では、市町村合併、学校区の検討、公共施設の利用圏域の設定などの、地域のまとまりを考慮すべき課題に知見を提供できると考えられる。

２．データベースの作成

(1) 研究対象

対象地域は、群馬県とする。群馬県は上毛三山（赤城山、榛名山、妙義山）、浅間山、谷川岳といった地域を象徴する多くの山々があり、山と触れ合う機会が多い地域である。分析対象テキストデータは歴史、文化が刻み込まれており、かつ県内に広く分布している中学校の校歌の歌詞とした。

(2) データベースの作成

本研究では、校歌の歌詞にテキストマイニングを適用するとともに、テキストデータと学校、地域の特性との関係を分析するため、表-1 に示すデータベースを作成した。群馬県内の公立中学校（市町村立）の 171 校を対象とした。

表-1　データベース

区　分		収集データ
特性	(1)学校	所在地・座標、旧市町村名、設立年、在校生数、学級数、校区内人口
	(2)立地	赤城山/榛名山/妙義山/谷川岳までの距離、学校の標高
	(3)望景	赤城山/榛名山/妙義山/谷川岳が見えるか・方角・山容、その他地物
テキスト	(1)校歌特性	制定年、作詞者名、作曲者名
	(2)校歌テキスト	校歌名、歌詞（1 番、2 番、3 番･･･）

３．地域イメージと地域のまとまりの把握

(1) 形態素解析

テキストデータを、何らかの方法で客観的・定量的に解析できるよう加工する必要がある。自然言語処理分野では、いくつかの解析手法やそのためのソフトウェアが開発されており、本研究では技術情報が公開されている KH Coder[5)6)] を使用することとした。本章で行う形態素解析とは、文や文章を、言語が意味を持つ最小単位の列に分割し、それぞれの品詞を判別する作業である。KH Coder では形態素解析に「茶筅（Cha Sen）[7)]」が組み込まれている。

(2) 語の出現頻度の集計

校歌テキストについて形態素解析を行い、助詞等を含め

表-2 語の出現頻度（地名、地物に関する語）

出現順位	抽出語	出現頻度	出現順位	抽出語	出現頻度
10	**赤城**	62	75	**榛名**	20
15	風	56	79	太田	19
16	山	54	82	館林	18
25	雲	41	100	桐生	15
28	道	39	124	**浅間**	13
36	みどり	35	125	富岡	13
44	高崎	31	126	利根	13
48	前橋	30	131	沼田	12
52	緑	29	138	笠懸	11
57	水	27	146	渡良瀬	11
61	川	25	147	**妙義**	11

青：地名に関する語、緑：地物に関する語

キーワード　イメージ、自然言語処理、テキストマイニング
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 TEL027-254-9179 E-mail：tmorita@cvl.gunma-ct.ac.jp

た全ての品詞の出現頻度を集計し、地名と地物に関する語を整理したものを表-2（前頁）に示す。地名に関する語は、山名である「赤城（10 位）」「榛名（75 位）」「浅間（124 位）」「妙義（147 位）」、市町村名等がみられる。地物に関する語は、「風（15 位）」「山（16 位）」「雲（25 位）」等がみられた。

(3) 地域のまとまりの分析

語の出現頻度の集計結果より、地域のまとまりを形成しているキーワードを「赤城」「榛名」「妙義」（以上、上毛三山）、「浅間」の山名を表わしていると考えられる 4 語とする。山頂から中学校までの距離帯別の語の出現率を図-1 に整理した。山頂に近いほど、その山名の語の出現率が高く、山頂を中心とした地域のまとまりがうかがわれる。図-2 は「赤城」の語が出現した中学校の分布をみたものであり、赤城山を中心とした地域が形成されていると考えられる。

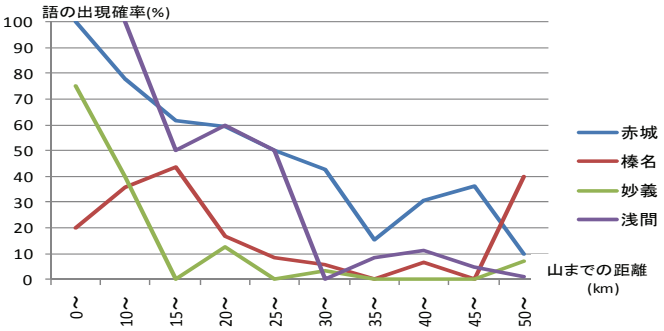

図-1 山頂から距離帯別の語の出現率

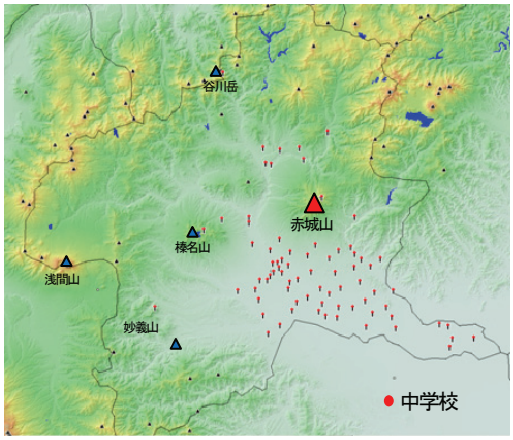

図-2 「赤城」が出現した中学校の分布

6．地域イメージの構成要素の分析

(1) 語の共起関係の算出

KH Coder は、形態素解析によって得られたキーワードについて、語と語の関連性を分析することができる。語と語の関連性を定量的に得る方法はいくつかあるが、比較的簡易な Jaccard 係数を用いることした。Jaccard 係数とは、語句 x と y がどれほど共起しているかを示す値である。値が大きくなるほど、語と語の関係（共起関係）が強いと考える。テキストデータ全ての集合を表-3 のように分割した上で、Jaccard 係数は次式で与えられる。

$$Jaccard(x,y)=\frac{a}{a+b+c} \qquad (1)$$

表-3 語句 x と y に対する 2×2 の分割表

	語句 y が現れる	語句 y が現れない
語句 x が現れる	*a*	*b*
語句 x が現れない	*c*	*d*

(2) 地域イメージの構成要素の分析

キーワードを中心に語の共起関係を分析し、地域イメージの構成要素を把握する。表-4 に、「赤城」に関する計算結果を示す。「赤城」は「高い」「明るい」と形容され、「嶺」「空」「風」「緑」「雲」との共起関係が高いことがわかる。

表-4 「赤城」に関する共起関係

順位	抽出語	品詞	Jaccard 係数	順位	抽出語	品詞	Jaccard 係数
1	希望	名詞	0.371	11	窓	名詞	0.247
2	中学校	名詞	0.343	12	風	名詞	0.237
3	光	名詞	0.315	13	学び	名詞	0.231
4	嶺	名詞	0.294	14	緑	名詞	0.225
5	理想	名詞	0.292	15	ゆる	動詞	0.220
6	仰ぐ	動詞	0.283	16	ぬ	助動詞	0.209
7	心	名詞	0.277	17	雲	名詞	0.205
8	高い	形容詞	0.265	18	輝く	動詞	0.202
9	空	名詞	0.260	19	若人	名詞	0.200
10	ある	動詞	0.256	20	明るい	形容詞	0.200

緑：地物に関する語、黄：形容詞

(3) 共起ネットワークの作成

共起関係を視覚的に表現する共起ネットワーク図を作成した。ネットワーク図は紙面の関係で講演会にて報告する。

5．まとめ

群馬県内の中学校の校歌の歌詞を対象にテキストマイニングを行い、地域イメージを構成するキーワードを抽出した。次に、キーワードによる地域のまとまりがあることを明らかにし、そのキーワードと他の語の共起関係を分析することにより、地域イメージの構成要素を把握した。

今後の研究課題は、1)学校特性、学校の立地特性、山の見え方等とキーワードとの関係を詳細に分析すること、2)共起関係について階層構造を分析すること、3)共起関係について語の意味を考慮した分析（Annotation）を行うことである。

謝辞：本分析には、KH Coder[5)6)]（著作権者、樋口耕一氏）を使用した。ここに記し、感謝の意を表します。

【参考文献】

1) 斉藤和夫・石崎裕幸・田村亨・桝谷有三：都市のイメージ構造と地域特性の関係に関する分析，土木学会土木計画学研究・論文集 Vol.14, pp.467-474，1997
2) 福田大輔・庭田美穂・屋井鉄雄：疑問型表現自由回答データを用いた社会資本整備に対する市民の関心の抽出方法に関する基礎的研究，土木学会土木計画学研究・論文集 Vol.24，pp.139-148，2007
3) 鄭蝦榮・羽鳥 剛史・小林潔司・白松俊：ファセット学習モデルを用いた公的討議のプロトコル分析，土木学会土木計画学研究発表会・講演集 Vol.36，2007
4) 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-，土木学会土木計画学研究・講演集 Vol.41，CD-ROM，2010
5) 樋口耕一：テキスト型データの計量的分析 —2 つのアプローチの峻別と統合—、理論と方法 19(1)、pp.101-115、2004
6) KH Coder，http://khc.sourceforge.net/，2011.1.19（閲覧）
7) ChaSen 形態素解析器，http://chasen-legacy.sourceforge.jp/，2011.1.19（閲覧）