地域の特性に基づく聞き書きの提示手法の提案

寺嶋　一将・植竹　俊文・竹野健夫（岩手県立大学大学院）

岩手県花巻市では地域の歴史や文化を後世へ残すことを目的に，郷土史研究団体が地域住民を対象とした聞き書きの収集を実施している．聞き書きはデジタルアーカイブ上で公開しているが，収集された聞き書きが 100 人分を超え，その分類・整理の手法が課題になっている．そこで，聞き書きに含まれる地域の特性を用いて収集された聞き書きを分類・整理する手法を提案する．本稿では地域の特性を求めるために文中の単語に対して特徴量を付与し，地域の特性を検討・評価した結果と，聞き書きを他の聞き書きや他史料と関連づけるための手法に関する実験結果を報告する．

Presentation Method of “Kikigaki” Using “Regional Characteristics”

Kazumasa Terashima / Toshifumi Uetake / Takeo Takeno (Iwate Prefectural University Graduate School)

Recently various organizations are engaging in preservation activity of local history and culture using digital archive. For example, local history research group in Hanamaki City, Iwate Prefecture records “Kikigaki” on digital archive. “Kikigaki” is a text collected by interview for citizens in Hanamaki City. Kikigaki is already collected from more than 100 citizrns. Collected "Kikigaki" is too much so we need to consider classification method. Therefore, we propose organize method of "Kikigaki" to use “Regional Characteristics”. In this paper, we report appraise method of "Regional Characteristics" and method of find the relation of "Kikigaki".

１．はじめに

近年，地域の歴史や文化を後世へ残すことを目的としたデジタルアーカイブの活用が全国的に広まっている[1]．岩手県花巻市ではこの活動の一環として，郷土史研究団体が聞き書きの収集を実施している．得られた聞き書きはデジタルアーカイブ上で公開[2]しているが，収集された聞き書きが 100 人分を超え，その分類・整理の手法が課題になっている．一般的なデジタルアーカイブが古文書や写真等，ある程度まとまった情報を管理しているのに対して，聞き書きは話者によって示される内容が大きく異なり，分類することが困難な状態にある．そこで，新たな整理の手法が必要になった．聞き書きは特定の地域に根付いた記録である．よって，多くの聞き書きには地域の特性が産業や文化等の形で文中に含まれていると考えられる．このことから，聞き書きに含まれる地域との関連性を用いて聞き書きを分類・整理する手法を提案する．本稿では，提案に用いる基本要素の方法論について報告する．

２．聞き書きの概要と現状

聞き書きとは話者の話を口述的な表現を活かしながら文章化した記録である．本稿では岩手県花巻市の住民を話者とした聞き書きを扱う．

２．１．聞き書きの概要

聞き書きは以下の図１の手順に従って収集される．取材の際に取材者が質問をすることもあるが，基本的に話者が過去に経験した出来事や生活，地域の伝統行事等について触れつつ，半生について自ら語る形式で取材は行われる．対象とする話者は郷土史研究団体に所属する人物の紹介等を通して集められる．また，できる限り高齢の人物から取材を行う方針が取られているため，大正時代に生まれた人物への取材が優先される．具体的な話者の男女比と出生年の分布を表１，表２に示す．取材は１～２時間程度行われ，得られた音声を書き上げる．さらに，この書き上げられた文章を共通する内容ごとにまとめ，４０００文字程度の文章に整えて公開される．

最終的な聞き書きの形式を図２に示す．聞き書きは話者の氏名や生年，略歴等の話者に関する基本的な情報が示された後に，話題ごとに小見出しが付与された複数の文章が記されている．まとめると，聞き書きは話者に関する基本的な情報と，複数の文章から構成されている．

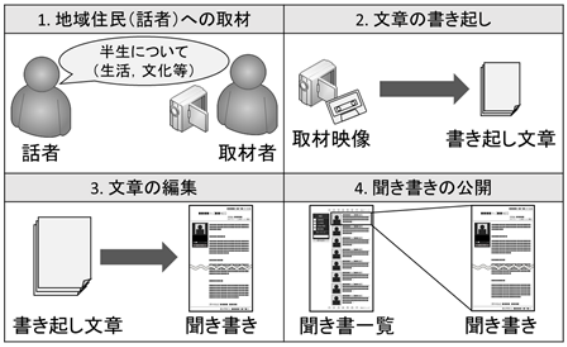

図 1　聞き書きの収集手順
Figure 1　Collection procedure of Kikigaki

表 1. 話者の男女比
Table 1　Gender ratio of speakers

	花巻	大迫	石鳥谷	東和	合計
男	28	22	14	13	77
女	12	6	4	2	24

表 2. 話者の出生年の分布
Table 2　Distribution of birth year

	花巻	大迫	石鳥谷	東和	合計
1910 年代	1	0	0	0	1
1920 年代	15	7	4	8	34
1930 年代	14	12	7	2	35
1940 年代	6	3	5	0	14
不明	4	6	2	5	17

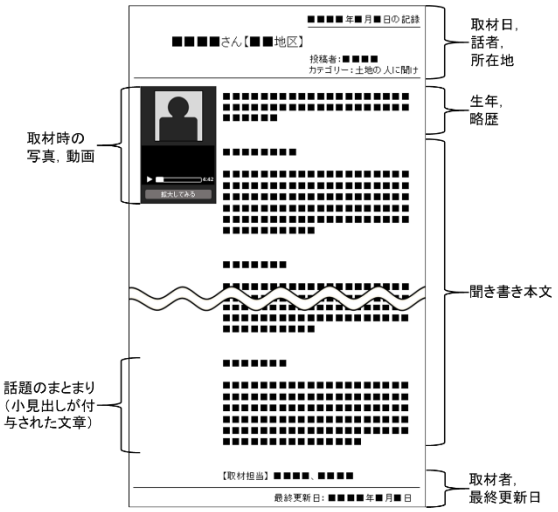

図 2　聞き書きの構成
Figure 2　Composition of Kikigaki

2．2．聞き書きの公開に関する現状

既に１００人を超える話者に対して取材を実施し，デジタルアーカイブ上で公開されている．聞き書きは現在，花巻市を合併前の４市町に分割したうえで，話者の所在地に従って地域ごとに検索できる形でまとめられている．しかし，現在の分類手法では，聞き書きの内容には触れていないため，聞き書きがもつ本質的な情報の発見・活用が困難な状態にある．また，聞き書きは話者の半生をまとめた記録であり，記される内容は話者によって大きく異なる．そこから，メタデータ等では整理することが難しい．

聞き書きは，基本的には地域で生きてきた人物の記録であり，地域に関する産業や文化，事件等の要素が複数の聞き書き間で共通することがある．本稿では聞き書きがもつこの特徴を利用して，地域の観点から聞き書きを分類・整理する．

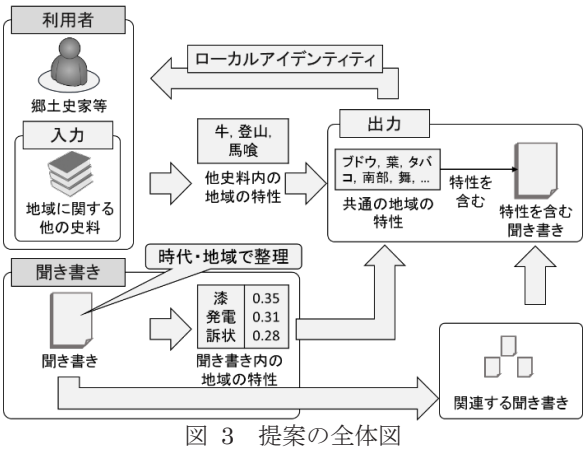

図 3　提案の全体図
Figure 3　Over view of proposal contents

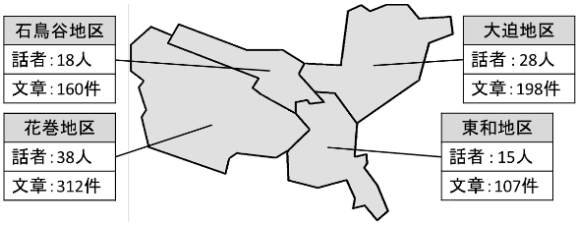

図 4. 各地域の話者数と文章数
Figure 4　Number of speakers and texts in each region

3．提案内容の概要

聞き書きの整理・分類手法として，『地域の特性』と『ローカルアイデンティティ』に着目する．それぞれの定義について，『地域の特性』は他の地域と比較した際に該当する地域で深く結びついている要素とする．これは語で表現できるものとして，その地域に関わる史跡や行事，人名等として現れるとする．『ローカルアイデンティティ』の定義に関しては，大堀ら[3]が示した解釈に従う．そこではローカルアイデンティティを『個人レベル』と『集合レベル』に分け，個人レベルを『個人の地域への帰属意識・愛着』，集合レベルを『地域の個性・らしさ』としている．本稿では『集合レベル』の定義に従う．また，ローカルアイデンティティは地域のキャッチフレーズ等の形で現れることも文献内で示されている．これにはキャッチフレーズが地域の将来像を示していることと，地域の歴史と乖離した将来像は破綻することが多いことが根拠として挙げられている．

それに合わせて，本稿ではローカルアイデンティティと地域の特性との関係をまとめて，「特定の地域において特性がもつ意味・役割を地域の個性として解釈したものをローカルアイデンティティ」と定義する．具体的には，「○○地区は，△△△△という人物が生まれた土地であり，その生家は今でも残っている」等の形で，各地域の特性がその地域内でどのように扱われているかが認識されたものとする．また，地域の特性及びロ

ーカルアイデンティティは一つの地域に複数存在するものとして扱う．

これらの定義を踏まえ，提案の全体図を図３に示す．最初に聞き書きを時代・地域の観点から分類し，語ごとに各地域においてどれだけ特徴的かを特徴量として数値で評価していく．その後，他の史料等に含まれる地域の特性と比較することで，他史料と共通する地域の特性を発見する．こうして得られた特性を含む文章と，その文章と関連した文章を抽出し，閲覧者に対して提示する．この際に，地域の特性と聞き書きをもとに閲覧者は地域像をローカルアイデンティティとして認識し，それを基に聞き書きを活用するしくみとする．

特徴量の計算や関連する文章の抽出などは聞き書き全文ではなく，小見出しが付与された文章単位で行う．また，対象とする地域の区分はデジタルアーカイブ上で現在用いられている，旧四市町に従う．地域区分の理由として，収集された聞き書きの件数が図４の分布になっていることが挙げられる．これは話者の数と小見出しが付与された文章の数をまとめた結果であるが，これ以上地域を細分化した場合に，地域の特性を発見するためのデータが不十分になる恐れがある．

３．地域・時代による分類手法

小見出しが付与された文章を内容に従って時代・地域の観点から分類する．地域の特性を求めるためには，文章を地域ごとに振り分けなければならない．合わせて，時代ごとに分類することによって，各時代による特性を導き出せると考えられる．そこで，これらの観点から文章を分類する手法について実験を行った．

３．１．地域による分類

話者は自身が住む地域に関する出来事等について語ることが多い．そこで，本稿では話者の所在地と文中で示される地域が基本的に一致すると仮定して，どの程度一致しているかを調査した．

調査を行ううえで，地名は旧４市町よりもさらに古い時代の行政区分を用いた．これらの地名の多くは現在でも使われているものである．調査の結果は図５である．図から，話者の所在地と文中に出現する地名が基本的に一致することが分かる．この結果から，聞き書きは話者の所在地に従って整理するものとした．

３．２．時代による分類

時代による分類は話者の生年等では難しいため，文中に出現する時代を示す語を用いて時代を推定する手法を考えた．この手法を用いた場合，文章全体を時代ごとに分類することは困難になるため，全文章の何割に時代を示す語が出現するかを調査した．調査した結果が表３である．対象とした語は和暦と西暦，「○○時代」の３種類の表記である．これらの表記の出現数と文章数から，表記の出現する割合を求めると，表４になる．各地域で全体の３割から４割程度の文章に時代表記が出現していることが分かる．対象とした文章の他に，時代の特定が可能な出来事等を時代表記として含めることで，時代の特定が可能な文章を増やすことも期待できる．ここから，時代の特定に関しては文中の時代表記を用いて整理するものとした．

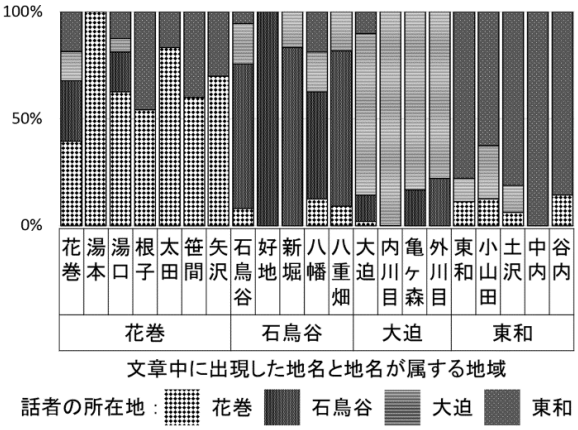

図 5. 話者の所在地と各地域の文中の地名出現率

Figure 5　Address of speakers and Place name written in texts

表 3. 各地域と時代における表記数

Figure 3　Number of era notation in each region and era

		地域				
		花巻	石鳥谷	大迫	東和	合計
時代	江戸時代以前	9	2	1	7	19
	明治～大正	19	10	9	9	47
	昭和	72	39	39	18	168
	平成	10	10	8	2	30
	合計	110	61	57	36	264

表 4. 全文章における時代表記を含む文章の割合

Figure 4　Percentage of sentences including era notation in all texts

	花巻	石鳥谷	大迫	東和	合計
各地域の文章数	312	160	198	107	777
時代表記を含む割合	35%	38%	29%	34%	34%

表 5. 現段階で定義が可能な各地域の特性
Figure 5 “Regional Characteristics” at the present stage

	特性	概要
花巻	宮沢賢治	花巻生まれの詩人・童話作家，農業の指導者としての側面も有名
	高村光太郎	花巻で晩年を過ごした詩人・彫刻家
石鳥谷	南部杜氏	日本酒の代表的な杜氏集団の一つ
大迫	早池峰山	北上高地の最高峰で貴重な高山植物が存在
	神楽	国の重要無形民俗文化財「早池峰神楽」が伝わる
	ワイン	ワインの産地で関連した催し物が開かれている
東和	萬鉄五郎	東和出身の画家で大正から昭和初期に活躍
	田瀬湖	田瀬ダムのダム湖でカヌー等の競技が行われる

表 6. 各特性を含む文書の出現数
Figure 6 Number of text including “Regional Characteristics”

	特性	出現文書数			
		花巻	石鳥谷	大迫	東和
花巻	(宮沢)賢治	22	3	9	−
	(高村)光太郎	22	−	−	−
石鳥谷	南部杜氏	−	5	−	−
大迫	早池峰	−	−	11	−
	神楽	4	3	18	3
	ワイン	−	−	4	−
東和	(萬)鉄五郎	−	−	−	1
	田瀬湖	1	−	−	5

表 7. 提案手法によって付与された特徴量
Figure 7 Feature amount by Proposed method

	特性	花巻	石鳥谷	大迫	東和
花巻	(宮沢)賢治	0.08	0.01	0.05	−
	(高村)光太郎	0.11	−	−	−
石鳥谷	南部杜氏	−	0.1	0.02	−
大迫	早池峰	−	−	0.06	−
	神楽	0.01	0.01	0.14	0.02
	ワイン	−	−	0.04	−
東和	(萬)鉄五郎	−	−	−	0.04
	田瀬湖	0.001	−	−	0.07

表 8. *TFIDF*によって付与された特徴量
Figure 8 Feature amount by TF-IDF

	特性	花巻	石鳥谷	大迫	東和
花巻	(宮沢)賢治	0.13	0.09	0.19	−
	(高村)光太郎	0.11	−	−	−
石鳥谷	南部杜氏	−	0.13	0.1	−
大迫	早池峰	−	−	0.06	−
	神楽	0.04	0.1	0.22	0.1
	ワイン	−	−	0.04	−
東和	(萬)鉄五郎	−	−	−	0.04
	田瀬湖	0.02	−	−	0.07

表 9. 残差 IDF によって付与された特徴量
Figure 9 Feature amount by RIDF

	個性	花巻	石鳥谷	大迫	東和
花巻	(宮沢)賢治	0.44	0.36	0.58	−
	(高村)光太郎	0.46	−	−	−
石鳥谷	南部杜氏	−	0.38	0.30	−
大迫	早池峰	−	−	0.29	−
	神楽	0.30	0.51	0.53	0.13
	ワイン	−	−	0.36	−
東和	(萬)鉄五郎	−	−	−	0.00
	田瀬湖	0.00	−	−	0.16

4．特徴量の付与に関する実験

地域ごとに文中の語がどれだけ地域において特徴的かを求める．これを特徴量として，文中の語に対して付与する式を(4)式として提案する．この式の有用性を調査するため，現時点で定義することができる地域の特性を用いて，既存の特徴的な語の評価に用いる式と比較した[4].

4．1．特徴量付与の提案式

(1)式は文中の語に特徴量を付与する TFIDF 法である．この値を基に各地域で出現しやすい語に対して重みを付与する(4)式を，地域の特性を発見するための手法として提案する．

文書T_jにおける単語W_iの特徴量を求める場合，(2)式の$n_{i,j}$は文書T_jにおけるW_iの出現数であり，$\sum_k n_{k,j}$はT_jにおける全単語の出現数の和である．(3)式のDは全文書数であり，d_iはW_iを含む文書数である．(4)式の$l_{i,m}$は任意の地域R_mにおけるW_iの出現する文書であり，L_iは全地域においてW_iの出現する文書数の和である．

$$TFIDF_{i,j} = TF_{i,j} \cdot IDF_i \quad (1)$$

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (2)$$

$$IDF_i = \log \frac{D}{d_i} \quad (3)$$

$$FV_{i,j} = TFIDF_{i,j} \cdot \frac{l_{i,m}}{L_i} \quad (4)$$

４．２．提案式に関する実験

提案式では各地域で出現しやすい語が地域の特性を発見する手掛かりになると仮定した．この仮定が正しいかを判断ために，一般的に知られている地域の特性が文中にどのように出現しているかの傾向を調べる．

まず，自治体のキャッチフレーズ等[5][6][7]から表５の語を地域の特性として定義した．これらの特性が，どの地域に住む話者の文章に出現したかを整理した結果が表６である．数値が記されていない箇所は特性が出現していないことを示す．また，下線が引かれた箇所は４地域で比較した際に，語が最も多く出現していた地域を示す．いずれの語も，特性として定義した地域で最も多く出現した．ここから，地域と特性の出現に関する仮定が正しいことが分かった．

次に，提案式を用いて定義した特性に対して特徴量を付与する．合わせて，既存手法との差異を調べるために一般的に，特徴的な語の評価に用いる，TFIDF 法と残差 IDF 法を用いて特徴量を計算する．(5)式は残差 IDF による特徴量を求める式である．E_iは名詞，nは全ての文書数，n_iは名詞E_iを含む文書数，F_iは名詞E_iの出現頻度である．

$$RIDF(E_i) = \log \frac{n}{n_i} + \log\left(1 - e^{-\frac{F_i}{n}}\right) \quad (5)$$

それぞれで付与された値を表７，表８，表９に示す．下線が引いてある箇所は４地域で比較した際に最も高い特徴量が付与された地域である．

結果から，既存手法が花巻地区の特性として挙がった「宮沢賢治」に対して，大迫地区で最も高い特徴量を付与したのに対して，提案手法では該当する花巻地区で最も高い値が付与された．大迫地区も宮沢賢治に関係する地域だが，今回の提案手法の趣旨は地域間を比較した際に他地域よりも強く特性を押し出している地域を見つけだすことにあるため，予め定めた地域と一致したことで，期待通りの結果が得られた．また，他の特性の値を比較すると，提案手法を用いることによって該当地域と他地域の値の差を拡大し，対象地域を際立たせることに成功していた．

また，各特性が地域内でどの程度，高い特徴量を得られたかを順位として表１０，表１１，表１２に示す．ここから，提案手法を用いることで順位に関しても向上する傾向を確認できる．また，地域間の差異も明確になっていることがわかる．

特徴量から得た順位を基に，特性の該当する地域における順位を手法ごとに比較した結果を表１３にまとめた．下線を引いた箇所は３手法を比較した際に最も高い順位を得た手法である．ここから，提案手法で最も多くの特性で他手法よりも高い順位を得られたことが読み取れる．

考察として，既存の特徴量付与の手法と比べて，提案手法を用いることによって他手法よりも優先的に地域の特性を際立たせることが可能と分かった．ここから，提案手法は地域の特性を発見する手がかりとして有用と考えらえる．

表 10. 提案手法による特性を示す語の出現順位
Figure 10　Appearance rank of proposed method

	特性	花巻	石鳥谷	大迫	東和
花巻	(宮沢)賢治	310	2195	446	–
	(高村)光太郎	148	–	–	–
石鳥谷	南部杜氏	–	102	1785	–
大迫	早池峰	–	–	281	–
	神楽	4908	2036	48	1429
	ワイン	–	–	602	–
東和	(萬)鉄五郎	–	–	–	351
	田瀬湖	5180	–	–	164

表 11. *TF-IDF* による特性を示す語の出現順位
Figure 11　Appearance rank of TF-IDF

	特性	花巻	石鳥谷	大迫	東和
花巻	(宮沢)賢治	173	293	35	–
	(高村)光太郎	267	–	–	–
石鳥谷	南部杜氏	–	124	246	–
大迫	早池峰	–	–	595	–
	神楽	1901	234	22	188
	ワイン	–	–	1248	–
東和	(萬)鉄五郎	–	–	–	783
	田瀬湖	4317	–	–	293

表 12. 残差 IDF による特性を示す語の出現順位
Figure 12　Appearance rank of RIDF

	特性	花巻	石鳥谷	大迫	東和
花巻	(宮沢)賢治	225	170	47	–
	(高村)光太郎	219	–	–	–
石鳥谷	南部杜氏	–	164	400	–
大迫	早池峰	–	–	438	–
	神楽	793	50	61	524
	ワイン	–	–	181	–
東和	(萬)鉄五郎	–	–	–	978
	田瀬湖	2321	–	–	508

表 13. 各手法による特性の出現順位
Figure 13　Appearance rank of each method

	特性	TF-IDF	残差IDF	提案手法
花巻	(宮沢)賢治	173	225	310
	(高村)光太郎	267	219	148
石鳥谷	南部杜氏	124	164	102
大迫	早池峰	595	438	281
	神楽	22	61	48
	ワイン	1248	181	602
東和	(萬)鉄五郎	783	978	351
	田瀬湖	293	508	164

５．聞き書きの関連付けに関する検証

聞き書きの関連付けとして，他史料との関連付けによる共通する地域の特性の発見と，聞き書き間の紐づけを想定している．そこで，提案手法を用いて他史料の地域の特性を評価し，聞き書きと比較した結果と，聞き書き同士を一般的に文章間の類似性の評価に用いられるコサイン類似度を用いて評価した結果を示す．

５．１．他史料との関連付けに関する実験

聞き書きから得られた特性を他史料と比較することを想定しているため，聞き書きとは異なる資料から同様に特性を抽出し，同一の地域で共通する特性として，どのような語が出現するかを調査した．

調査に用いたのは，デジタルアーカイブ内で聞き書きとは別コンテンツとして公開されている花巻市で発行された新聞である．公開される記事は地域の特別な行事や大きな出来事について中心的に取り上げられているものである．今回は記事の数を考慮してうえで，１９５５年の新聞記事２２１件を対象とした．合わせて，時期の観点からも特性に変化が起こり得るかを調査するために，３カ月ごとに記事を分け，それぞれの期間に対して提案手法を用いて特徴量の計算を行った．地域と時期に基づく記事の分布を表１４に示す．

表 14. 各期間と地域における記事の数
Figure 14　Number of articles in each period and region

地域	1月～3月	4月～6月	7月～9月	10月～12月	地区合計
花巻	34	34	32	31	131
石鳥谷	8	8	8	4	28
大迫	4	7	8	4	23
東和	14	14	5	6	39
期間合計	60	63	53	45	221

表 15. 記事と聞き書きの双方で高い特徴量になった語
Figure 15　Words with high feature amount on both sides

	1月～3月	4月～6月	7月～9月	10月～12月
花巻	牛,発明,宮沢,旅館,矢沢,花巻温泉,郎,文化,発表,展,山口,母,姉,候補,達,内容,会員	桜,移転,沼,電車,笹,合併,豊沢,内容,開設,事件,温泉,候補	賢治,駅,イギリス,海岸,通信,日本,堤防,川,高村,島,氏,電鉄,字	堤防,重次郎,宮沢,観音,事件,基地,アイヌ,間,温泉,猫,宮野,鯉,寸,催し,結核,氏
石鳥谷	八重畑,石鳥谷		様式,制,参加,農業	八重畑
大迫	助役,役,収入	ブドウ,葉,タバコ,南部,舞,食肉,集団	池峰,あんど,登山,祭り,大迫	神楽,奉納,山伏,外川目,ブドウ
東和	紙,ダム,田瀬,山,土沢,収入,東和	鮎,発電,胆沢,東和,釣り,湖,田瀬,ワカサギ	発電,級,田瀬	浮田,東和,田瀬

聞き書きと新聞記事の双方に現れた特性の判断として，聞き書きに対しては，地域ごとに上位１０００位以内の語を「高い特徴量をもつ語」とした．新聞記事に対しては，聞き書きよりも語の数が少ないことから，順位ではなく特徴量で制限を設けることとして，特徴量の値が０．０５以上の語を「高い特徴量をもつ語」とした．双方において高い特徴量をもつ語に該当するものを表１５に示す．下線を引いた語は地域と深く結びついていると判断できる語である．

結果及び考察として，地域と深く結びついている語には地名も多く含まれているが，その土地の地形や産業，文化に係る語が多数見られた．聞き書きのみを対象として特徴量を付与した場合よりも地域の特性を絞り込むことができたことから，他の史料と併用した地域の特性の評価への活用が期待できる．また，地域の特性を用いて他の史料と関連付けられることを確認した．

５．２．聞き書き同士の関連付けに関する実験

地域の特性だけではなく，異なる観点から文章の関連性を求めることで，提示する情報の網羅性を高める．そこで，文章間の類似性を調べる代表的な方法論として，コサイン類似度に着目した．この手法を用いて文章の中から類似する組み合わせを発見し，そこから新たな知見等を得られるかを調査した．

大迫地区の聞き書きに限定して，コサイン類似度の計算を行った．対象は話者が２８人で，文章に分解すると１９８件になる．まず，各文章と残りの１９７件の文章を比較して，各文章と最も類似の高い文章の組を求めた．得られた１９８組を，類似度の高い順に並べ替えた．その結果が表１６である．ここでは類似度の範囲を６段階に分けて，各範囲において上位の組を最大５件抽出し，それらの組の類似性を文章の内容で確認した．

表 16. 類似する文章の小見出し（一部抜粋）
Figure 16　Title of similar texts

類似度の範囲	比較元	類似する文章
0.60 以上	馬産地 (A)	カイコと馬 (B)
	神楽があったからこそ	神楽とは
	馬産地だった (C)	馬産地 (A)
0.50 ～ 0.59	大迫と亀ヶ森	役場の仕事
	タバコ栽培	タバコづくり
	タバコ栽培 ※上と別話者	タバコづくり
	家で行った結婚式	役場の仕事
	馬の話 (D)	馬産地 (A)
0.40 ～ 0.49	祭り	大迫の秋祭り
	内川目の風習「馬っこつなぎ」(E)	馬産地 (A)
	古文書と年貢地頭	大迫と亀ヶ森
	昔は風流だった	芸者さんを追う子供たち、お茶道具を売る店
	馬と蚕 (F)	カイコと馬 (B)
0.30 ～ 0.39	悲惨な空襲体験、兄はアッツ島で玉砕	役場の仕事
	農地改革、教職を辞めて農業に従事	稲刈りと田打ち
	早池峰山	役場の仕事
	集落に伝わる行事「馬っこつなぎ」の馬のわら人形づくり (G)	馬産地 (A)
	雪と寒さと水	昔との変化
0.20 ～ 0.29	盛んだった炭焼き	大迫と亀ヶ森
	弁当のおかずを自作	戦後の食べ物の思い出
	戦時中の様子	戦争中の思い出
	けん葉	乞食のこと
	養蚕と久留米絣	森の幸と桑の実
0.10 未満	ナタネ油とランプ生活	松ヤニを焚いて灯りに

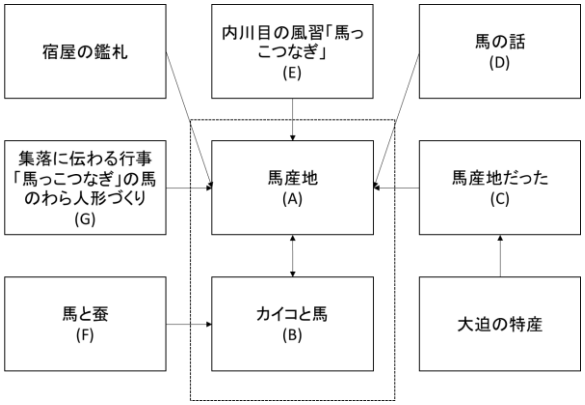

図 6. 文章の関連図
Figure 6　Association chart of Kikigaki

その結果，上位の組ほどより具体的な内容を共有していることが確認できた．表１６の，タバコに関する文章であれば，品種やブランド名等の細かな要素が一致している傾向が見られた．しかし，類似性の薄い文章が複数個所で出現する傾向も見られた．「役場の仕事」は４か所で出現しているが，他の文章に比べると明らかに比較元と異なる内容を示していた．ここから，おおよその文章では類似する他の文章を発見することが可能だが，ノイズになる文章が一部含まれているため，それらへの対策が必要であることが分かった．

類似する文章を俯瞰して観察すると，馬に関する記述が多く見られた．そこで，最も高い類似度を得た「馬産地」と「カイコと馬」の２件の文章を中心として，類似する文章を整理した．その結果が図６である．矢印は比較元から類似する文章に対して伸びている．表１６に出現する小見出しと共通する小見出しには(A)～(G)の同じアルファベットを付与している．また，図６では省略しているが「大迫の特産」は６件の文章と紐づけられていた．それ以外の図６の文章は図中に記した関係以外に他の文章と繋がることは無かった．

結果として，この図をもとに本文の内容を見ると「馬の競り」や「他産業との兼業」，「伝統行事」等，馬を中心とした地域に関わる事象で複数の文章を紐づけることができていた．特に他産業との関わりについては養蚕やタバコ等，当時は馬の飼育だけではなく，他の仕事も並行して行っていたことが読み取れた．ここから，単体の文章では一事例に過ぎない事柄が，複数の文章と併せてみることによって新たな知見として示すことできた．コサイン類似度を用いて文章を関連付けることで，新たな地域像の発見等への活用が期待できる．

図６では最も強く類似している文章同士で繋げたが，２番目以降に類似している文章等を含めて，対象を広げることで，より多くの文章を紐づけることが可能になると考えられる．同時に関係の薄い文章が含まれるリスクも増えるため，今後

は類似度に閾値を設ける手法について実験を進めていく．

６．おわりに

本稿では聞き書きの提示手法を提案し，そこで用いる基本になる３種類の方法論について実験と評価，考察を行った．それぞれ，時代・地域の観点からの聞き書きの分類，地域においてどれだけ特徴的かを表す特徴量の付与，聞き書きの関連付けの３種類について報告したが，どの手法についても提案する聞き書きの提示手法への活用が見込める結果が得られた．

聞き書きの分類については地域の観点について，話者の所在地とおおよそ一致する傾向が明確に表れたため，文章全体を地域ごとに分類することが可能になった．課題としては時代による分類が可能な文章を増やす必要がある．全ての文章を整理することは困難だが，時代を特定する出来事等を調査し，時代表記として追加することで，より多くの文章の分類が可能になると考えられる．特徴量の付与については，提案手法を用いることによって，あらかじめ定義した地域の特性に対して，該当する地域において高い特徴量を付与することが可能になった．聞き書きの関連付けについては，地域の特性に基づく史料間の関連付け等への活用が可能と考察できた．今回は新聞記事を用いた事例を報告したが，今後は他の史料を用いた場合についても実験を行い，史料の種類による結果の変動について調査する．コサイン類似度を用いた文章同士の関連付けについては，地域像の提示へ活用が期待できた．

今後は，時代表記の追加やコサイン類似度の閾値等の課題の解消に向けた追加実験等を行う．それに合わせて，本稿で取り上げた基礎的な要素の実験・評価結果に基づいてシステムの実装と評価を実施する．

参考文献

[1] 総務省関東総合通信局情報通信連携推進課：地域住民参加型デジタルアーカイブの推進に関する調査検討会報告書(2010).
[2] ふるさと遺産研究所：花巻物語辞典, <https://hana-isan.com/Home>(参照 2018-04-01).
[3] 大堀 研：自治体戦略としての「ローカル・アイデンティティの再構築」，社会学年報 Vol. 40，pp.23-33(2011).
[4]寺嶋一将，植竹俊文，竹野健夫：郷土史料を用いた地域の特性の抽出手法，―ききがきを活用したローカルアイデンティティの発見―，情報文化学会誌, Vol.25, No.1, pp.35-42 (2018).
[5]花巻市：宿場町おおはさま４００年記念事業４００年を迎える神楽とワインの里・大迫(2017).
[6]花巻市：第 37 回南部杜氏の里まつり，花巻市(オンライン)，入手先 <https://www.city.hanamaki.iwate.jp/event/1601/p007893.html>(参照日 2017-07-02).
[7]花巻市：花巻市消防本部管内の概況，平成 27 年版花巻市消防年報，p.1(2015).