2ZH-04

地域に関する史料を用いた聞き書きの活用モデルの提案

寺嶋　一将†　　植竹　俊文†　　竹野　健夫†

岩県立大学大学院　ソフトウェア情報学研究科†

1. はじめに

岩手県花巻市では郷土史研究団体が市民の半生の資料的価値に着目して，取材による調査を行ったうえで，聞き書きとしてデジタルアーカイブ上で公開している[1]．収集された聞き書きの増加に伴い，その新たな活用が模索されている．

市民の目線から得られた史料である聞き書きは既存の史料とは異なる観点から地域像を探る手がかりになる．よって，聞き書きを他史料と併用した地域の研究・学習への活用が期待できる．以上から，花巻市に関する新聞記事を用いて，聞き書きと新聞記事から得られる地域像を比較調査した．その結果から地域の研究・学習で聞き書きを活用するモデルについて考察する．

2. 聞き書きの概要

聞き書きとは古老等の話を語り口調を活かしつつ書き起した文章であり，話者の半生を通して地域の歴史や文化等について言及される．

2.1 聞き書きの概要

現在，聞き書きは現花巻市を花巻地区，大迫地区，石鳥谷地区，東和地区の 4 地域に分けて，話者の所在地に従って分類されている．それぞれの地域には産業や文化，地形等に独自の特性があり，聞き書きでは話者の半生に関連させて，その特性に言及されることが多い．聞き書きはそれらの特性について説明した話題ごとに文章がまとめられることが多く，各文章に小見出しが付与されている(図 1)．話者の数と小見出しが付与された文章の数を表 1 に示す．

2.2 聞き書きの現状・課題

2019 年 1 月現在，123 名の話者の聞き書きが公開されている．取材は継続して実施中であり，今後も公開される聞き書きの数は増加が見込まれる．収集量の増加に伴い，聞き書きの新たな活用方法が模索されている．そこで，他史料と聞き書きの連携した活用が考えられる．しかし，聞き書きは近代の多様な事柄を扱った文章群であり，内容の多様性から他史料との連携が難しい状態にあることが課題になっている．

Proposal of usage model to use “Kikigaki” with historical materials

†Kazumasa Terashima, Toshifumi Uetake, Takeo Takeno
Iwate Prefectural University Graduate School of Software and Information Science

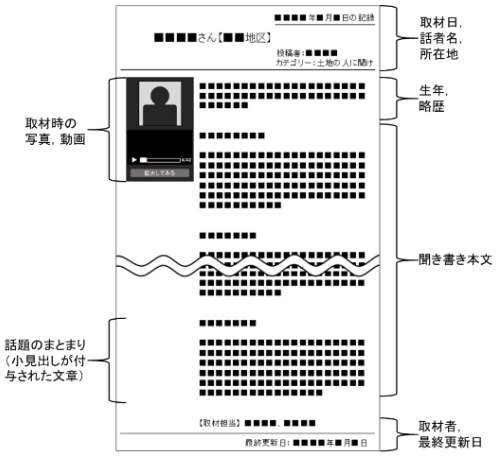

図 1.　聞き書きの構成

表 1.　聞き書きの話者数と文章数

	花巻	石鳥谷	大迫	東和	合計
話者	38	18	28	15	99
文章	312	160	198	107	777

3. 提案モデルの概要と検証

聞き書きと地域に関する他史料の内容から地域の特性を発見する．それに従って聞き書きと他史料を紐づけて提示することにより，地域の研究・学習に活用するモデルを提案する(図 2)．

3.1 検証の概要

聞き書きがもつ地域の特性に着目して，地域の特性で聞き書きと他史料を関連付ける．そこで，聞き書きと地域について記した新聞記事の内容に従って地域の特性を評価・抽出する．記事は聞き書きと同じデジタルアーカイブで公開している史料である．記事と聞き書きの文章から各地域の特性を評価・比較して，共通する地域の特性を調査する．

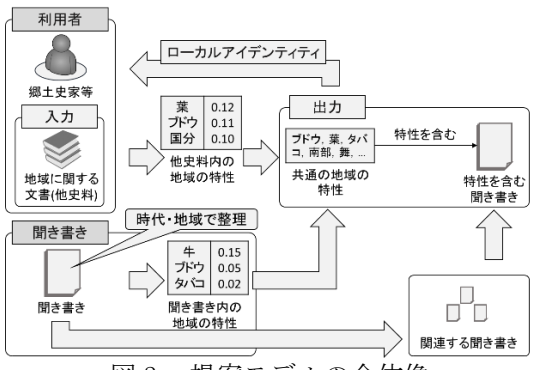

図 2.　提案モデルの全体像

3.2 地域の特性の評価手法

話者は所在地周辺の事柄を中心に語る傾向があり，話者の所在地と聞き書きの内容が示す地域は一致する傾向にある．また，地域の特性になりうる語はその地域の文章内に他地域よりも現れやすい傾向がある．以上より，TFIDF 法と，地域における語の出現頻度を用いて，文中の語が各地域にとってどれだけ特徴的かを特徴量として評価する式を(4)式として定義する[2]．

文書T_jにおける名詞W_iの$TFIDF$は(1)式で表される．$n_{i,j}$は文書T_jにおける単語W_iの出現回数であり，$\sum_k n_{k,j}$は文書T_jにおける全単語の出現回数の総和である．Dは全文書数であり，d_iは単語W_iを含む文書数である．提案式(4)の$FV_{i,j}$は求める特徴量である．$l_{i,m}$は任意の地域R_mにおける単語W_iの出現する文書であり，L_iは全地域における単語W_iの出現する文書数の総和である．

$$TFIDF_{i,j} = TF_{i,j} * IDF_i \tag{1}$$

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

$$IDF_i = \log \frac{D}{d_i} \tag{3}$$

$$FV_{i,j} = TFIDF_{i,j} * \frac{l_{i,m}}{L_i} \tag{4}$$

3.3 特徴量の付与と比較

聞き書きの文章（内容のまとまり）と新聞記事に含まれる語に特徴量を付与した．新聞記事は 1955 年の記事を対象として，季節により結果に差が現れる可能性を考えて，4 つの期間に分けて，合計 16 の文章のまとまりごとに特徴量を計算した(図 3)．新聞記事の数と期間・時期による分布を表 2 に示す．

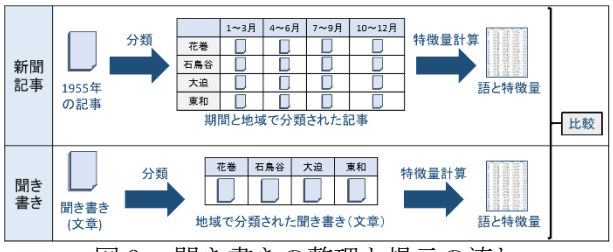

図 3　聞き書きの整理と提示の流れ

表 2　期間と地域による記事の分布

	1 月～3 月	4 月～6 月	7 月～9 月	10 月～12 月	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
合計	60	63	53	45	221

表 3　記事と聞き書きの双方で高い特徴量の語

	1 月～3 月	4 月～6 月	7 月～9 月	10月～11 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

3.4 検証結果・考察

聞き書きと新聞記事の双方に現れた地域の特性に関して，聞き書きに対しては各地域で特徴量が上位 1000 位以内の語，新聞記事に対しては特徴量が 0.05 以上の語を「高い特徴量をもつ語」と定義した．記事と聞き書きの双方で高い特徴量を持つ語を表 3 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．

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．聞き書きや記事のみを対象として特徴量を付与した場合には上位に地域の関係を読み取りにくい語も多数現れたが，史料の併用によって地域の特性を絞り込むことができた．以上から，聞き書きと他史料を併用することで，明確な地域の特性の発見が可能なことを確認した．

4.　おわりに

検証により聞き書きと他史料を併用することで地域の特性を評価・発見することが可能と分かった．以上から，聞き書きの活用モデルにおいて，他史料と聞き書きを地域の特性の観点から紐づけて示すことが可能と考察できた．また，聞き書きの提示モデルで地域の特性を用いることで，他史料との連携に合わせて，利用者に対する新たな調査の観点の提供が期待できる．

参考文献

[1] ふるさと遺産研究所：花巻物語辞典, <https://hana-isan.com/Home>(参照 2019-01-08).

[2] 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).