特定地域に限定しない観光キーワードの自動抽出

遠藤　雅樹[†,††]　横山　昌平[†††]　大野　成義[†]　石川　博[††]

† 職業能力開発総合大学校　〒 187–0035 東京都小平市小川西町 2–32–1
†† 首都大学東京　〒 191–0065 東京都日野市旭が丘 6–6
††† 静岡大学　〒 432–8011 浜松市中区城北 3–5–1
E-mail: †{endou,ohno}@uitec.ac.jp, ††yokoyama@inf.shizuoka.ac.jp, †††ishikawa-hiroshi@tmu.ac.jp

あらまし　我々は，Web 上の観光情報に着目し，旅行者が必要とする地域の観光情報を自動抽出し同質の情報を融合する手法を検討している．一般的に，辞書登録のない未知語は，人手による辞書登録などコストをかけて管理・運用することが多いが，我々は，人手によりコストをかけて生成する辞書は利用せず，形態素 N-gram と $RIDF$ による重み付けを利用して，地域サイトの情報から対象地域の観光キーワードを自動取得する手法を提案している．本稿では，複数の地域サイトから辞書登録に関係なく観光スポットなどの観光キーワードを自動取得した結果を示し，本手法を使うことにより特定地域に限定せず低コストに対象地域の有用な観光キーワードを取得できることを報告する．

キーワード　観光情報，キーワード抽出，Web マイニング

1. はじめに

近年，Web 上に存在する大量の情報から Web 利用者の必要とする有用な情報を的確に取得する技術の確立が求められている．高速通信網の普及や通信デバイスの充実に伴い，様々な種類の情報が大量に Web 上に置かれる傾向にある．そして，それらの情報は日々刻々と増加しており，今後もあらゆる分野の情報が Web 上に置かれることが予想されている．これに伴い，Web 利用者が必要とする有用な情報を取得することが困難になっている．

ここで，Web 上の情報を観光情報に限定する．観光情報には，国や地方自治体，企業の発信する地域サイトからブログや Twitter などを利用した個人の投稿まで多種の情報が提供されており，大量の情報が点在している．このため，Web を利用して旅行者が旅行計画時や旅先で観光情報検索を行い，有用な観光情報を取得することは一般的な作業となっている．

しかし，Web 利用者が必要とする有用な観光情報を取得することは，Web 上に点在する観光情報の増加と共に困難になっている．これは，検索エンジンなどから得られた検索結果から観光情報を抽出し，関連した観光情報を融合する処理が，システム化されていないためである．ここでの，情報抽出 (抽出) は，検索結果から取得した Web テキストが観光情報か否かを区別し，どの観光スポットについて記述されているかを判断することを指す．また，情報融合 (融合) は，関連した観光スポットについて記述された Web テキスト同士を結び付けて観光に役立つ情報とする作業を指す．観光情報の抽出とその融合は，システム化されておらず人手で行う作業であるため，Web 利用者個々の情報処理能力に依存している．このため，情報弱者にとっては，検索結果から観光情報を抽出し，観光スポットや観光スポットに関連するロコミ情報を融合して有用な観光情報を得る作業を繰り返すことは容易ではない． 仮に検索結果が数十件であったとしても，その情報を人手によって分析し融合する作業にはコストがかかる．

そこで，我々は，Web 上に存在する大量の情報から有用な観光情報を自動抽出・融合する手法として，形態素と N-gram 組み合せたキーワード抽出と $RIDF$($ResidualIDF$; 残差 IDF) を利用した重み付けによる観光キーワードの抽出手法を提案している [1]．特に，本研究では，人手による辞書作成など頻繁なメンテナンスを必要せず低コストに運用可能な点と，特定地域に限定せず観光情報を自動抽出・融合可能な点に着目した手法を議論している．

本稿では，複数の地域サイトから辞書登録に関係なく観光スポットなどの観光キーワードを自動取得した結果を示し，本手法を利用することで，特定地域に限定せず低コストに対象地域の有用な観光キーワードを取得できることを報告する．

2. 関連研究

観光情報に関連する研究は，Web における観光情報の提供や分析など，様々な研究やシステム開発が行われている [2]．

本研究は，Web 上の観光情報を抽出・融合し提供することを目的としているため，類似した研究として，2 種類の関連研究を挙げることができる．

1 つは，観光 Web コンテンツの分析による情報発信状況の抽出である．

三田村ら [3] は，ブログを収集しブログマイニングを行うことで，観光キーワードの出現頻度を調査し，観光とブログとの関係に着目し検討を行っている．

守屋ら [4] は，図書館情報におけるシソーラス構築のノウハウを応用して，観光に特化した観光情報シソーラスの設計の可能性について研究している．

石野ら [5] は，旅行ブログエントリから自動的に土産情報と観光名所情報を抽出する手法を提案している．

寺西ら [6] は，観光ガイドブックのページをカテゴリに分類することで構造化し，旅行ブログエントリと質疑応答コンテン

ツの対応付けを行っている．

これらは，観光キーワードを抽出する点では非常に類似しているが，我々は，特定のサイトではなく，地域サイトやブログなど異質で多様な Web 文書を対象に抽出を行う点に特徴がある．また，人手による辞書作成などの事前準備に頼らず低コストに観光キーワードの抽出を行い，自動抽出した観光キーワードを基準に抽出・融合処理を行う点にも特徴がある．地域サイトの情報更新に合わせて観光キーワードを自動で更新する機能を備えることで機械学習により抽出・融合可能であり，低コストに最新の観光情報を収集可能である．このため，人手によるコストを必要としない特定地域の観光情報抽出を行うことが可能であり，これまで研究されてきた人手をかけて辞書を充実させて精度を向上させる観光情報に関する研究とは手法が異なる点に特徴を持つ．さらに，本研究では，特定分野の辞書を利用する情報抽出・融合手法ではないため，観光情報以外の情報抽出・融合にも適用できる可能性があり，Web 上の大量の情報から真に必要とする情報を取り出す技術となる点にも特徴がある．

2 つ目は，観光イベントなどの情報検索技術である．

森本ら [7] は，事前登録のない施設情報を自動抽出する検索システムを開発している．これは，施設の種別などのキーワードと地名の一部を利用し，Google Maps API を応用したロボット型施設検索システムである．

小作ら [9] は，新聞記事コーパスでの単語出現特徴から観光イベント情報の検索支援を行っている．

これらは，観光情報を検索する点で非常に類似している．また，事前に登録されていない施設や観光イベント情報を検索する点も我々と目的は同等である．しかし，我々は，Web 上に点在する施設やイベント情報などの観光情報だけでなく，観光情報に関連した口コミ情報など周辺情報も融合し取得する点に特徴がある．提案手法は，Web 上に点在する多量の情報から同質の情報を関係付けて融合する新たな手法となる可能性がある．これは，観光イベントなどの観光キーワードに関する情報に対して付加価値のある口コミ情報などを融合させることができるため，より有用な情報として扱うことができる．

3. 提案システムの概要

本章では，Web を利用して観光情報の自動抽出・融合を行うシステムについて記述する．3.1 節に本研究におけるシステム利用者の想定，3.2 節に提案システムの概要を記述する．

3.1 システム利用者の想定

我々は，Web を利用して観光情報を検索するユーザを想定した観光情報の自動抽出・融合手法の提案する．

一般的に人が旅行をする際，事前の旅行計画検討や旅行先での情報収集など，必要に応じ観光情報の収集を行う．このときの Web を利用した情報収集を我々は予め次の手順で行うことを想定した．

まず，観光地の地名や名称から既存の検索サイトを利用し，観光地の情報について検索する．検索結果から，自治体や観光協会の提供するポータルサイトや旅行代理店などのサイトを参照し，観光地の観光スポットやアクセス方法などの詳細情報を取得する．この検索を繰り返し，観光地周辺の複数の観光名所や施設の詳細情報を収集する．しかし，詳細情報を取得するだけでは意思決定するための情報としては不足している．

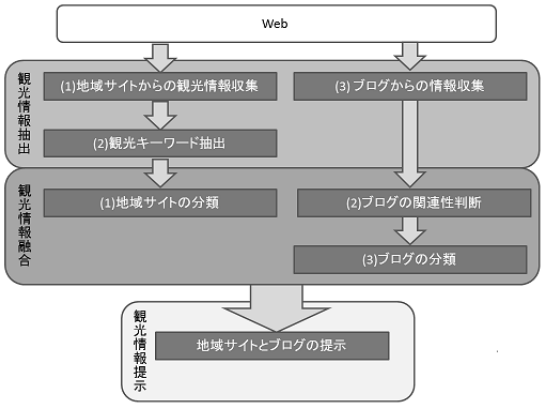

図 1　システム処理手順
Fig. 1　System processing procedure.

そのため，次に，取得した観光名所や施設に関して他の旅行者などが紹介しているロコミや評判などを検索する．名所や施設の他者の評価情報を取得し，地域サイトなどで取得した詳細情報と関連付けて参考にすることで意思決定の参考となる観光情報として利用している．

このように，Web を利用した観光情報の収集は，地域サイトなどから観光スポットなどの詳細情報とブログなどから観光スポットの評判情報を検索する作業を繰り返しながら行う．何度も検索を行うことで多くの観光スポットについての詳細情報や評判情報を取得可能であるが，一方で手間や時間など多くのコストがかかる．また，取得した詳細情報と評判情報は，関連付ける作業を人手で行うことで観光する場所を決める意思決定の材料となる．しかし，この作業は，検索者の情報検索・融合能力に大きく依存しているため情報弱者には困難な作業である．現在，Web 利用者が観光の意思決定に有用となる観光情報を自動で抽出・融合する作業はシステム化されていない．

そこで，本研究では，Web 上から観光情報を抽出し関連する情報を融合する作業を自動化するシステムを提案する．このシステムにより，利用者の情報検索・融合能力に依存せず，大量の情報から有用な観光情報を自動で取得する手法の実現に結び付けたいと考えている．

3.2 提案システムの概要

我々は，Web 上から検索対象とした観光地の観光情報を自動抽出・融合する作業をシステム化する手法の提案を行った [8]．提案手法は，大きく観光情報抽出と観光情報融合及び観光情報提示の 3 つの機能で構成される．図 1 に，提案システムの処理手順を示す．3.2.1 項に観光情報抽出，3.2.2 項に観光情報融合，3.2.3 項に観光情報提示の概要を示す．

3.2.1 観光情報抽出の概要

観光情報抽出は，地域サイトからの観光情報収集と観光キーワード抽出及びブログから情報収集として次の処理を行う．

（1） 地域サイトからの観光情報収集

表 1　カテゴリと事前定義語
Table 1　Category and pre-defined words.

カテゴリ	事前定義語		
見る・遊ぶ	観光スポット	見どころ	遊ぶ
イベント・祭り	祭り	イベント	催し
食べる・泊まる	グルメ	宿泊	味覚
お土産・特産品	土産	特産	工芸
自然・文化	景勝地	歴史	史跡

（2）　観光キーワード抽出
（3）　ブログからの情報収集

(1) 地域サイトからの観光情報収集は，観光地のある自治体や観光協会の提供する地域サイトから観光地の名所や施設についての概要やアクセス方法などの観光情報を収集する．一般的に自治体や観光協会の提供する情報は信頼性が高いため，地域の観光名所や施設などの情報の多くを収集できると考え収集対象とした．

(2) 観光キーワード抽出は，収集した観光名所や施設について記述された地域サイトの文書を利用し，検索対象地域の観光を表す観光キーワードの抽出を行う．ここで，観光キーワードとは，観光名所や施設及びイベントの名称とその名称に関連のある特徴語を指す．観光キーワードは，形態素解析器 (Mecab[注1]) による形態素解析結果から形態素と出現した順序を基に形態素 N-gram を取り出し，$RIDF$ を基準とした重み付けにより抽出した．ここで，Mecab の辞書は，標準の IPA 辞書を利用し人手による辞書登録などは行わないものとしている．この処理により地域サイトから抽出した観光キーワードを基に 3.2.2 項の観光情報融合を行うこととした．観光キーワード抽出については，4 章で詳細を述べる．

(3) ブログからの情報収集は，検索対象とした地域名を基に地域名が文書内に使用されているブログページを収集する．

3.2.2　観光情報融合の概要

観光情報融合では，地域サイトの分類とブログの関連性判断及び分類を行う．この処理は，3.2.1 項で記述した観光情報抽出により取得した地域ページとブログページ及び地域サイトから抽出した観光キーワードを利用する．

ここで，観光情報融合の前処理として分類の基準となる 5 つのカテゴリを定めた．5 つのカテゴリは，富山県内の 15 市町村の地域サイトのカテゴリを参考に,「見る・遊ぶ」,「イベント・祭り」,「食べる・泊まる」,「お土産・特産品」,「自然・文化」とした．また，各カテゴリには，分類の基準となるキーワードとしてカテゴリに含まれると考えられる事前定義語を予め人手で用意した．事前定義語は，5 つのカテゴリを合わせて約 60 語とした．表 1 に，各カテゴリと事前定義語の例を示す．

観光情報融合は，3.2.1 項の観光情報抽出で地域サイトから抽出した観光キーワードを利用し，次の処理を行う．

（1）　地域サイトの分類
（2）　ブログの関連性判断
（3）　ブログの分類

(1) 地域サイトの分類は，地域サイトの各ページを 5 つのカテゴリに分類する．収集した地域サイトの各ページが表 1 に示す 5 つのカテゴリのどのカテゴリに属するかを，事前定義語として予め用意した約 60 語の出現頻度を基準に排他的に分類を行う．ここで，事前定義語に含まれる語が存在しない地域サイトのページは，非分類となりカテゴリに属さないため以降の処理では利用されないこととなる．

(2) ブログの関連性判断は，ブログの各ページが検索対象地域の観光情報に関連があるページかを判断する．ここでは，線形判別関数を利用した．線形判別関数は，収集したブログページから学習データとして 50 ページを利用し各ページの観光キーワードの出現回数を基準に生成した．関連性判断により検索対象地域の観光情報に関連があると判断したブログページを，処理 (3) ブログの分類で利用する．

(3) ブログの分類は，関連性判断により関連性があると判断したブログページを 5 つのカテゴリに分類する．分類には，処理 (1) によって 5 つのカテゴリに分類された地域サイトページ群に含まれる観光キーワードを利用した．各ブログページでの観光キーワードの出現頻度を求め出現頻度の多いカテゴリに分類する．地域サイトの分類では，表 1 に示した事前定義語を基に分類を行うが，ブログの分類には事前定義語は用いていない．5 つのカテゴリに分類された観光キーワードを利用することで，事前定義語を利用した場合と比べ，より正確に地域サイトページ群とブログを関連付けることができる．

3.2.3　観光情報提示の概要

観光情報提示では，観光情報抽出と観光情報融合によりカテゴリに分類された地域サイトとブログの情報をシステム利用者に提示する．現段階では，カテゴリを選択すると観光キーワードに関連付けられた地域サイトとブログのページのリンクを提示する．

4.　観光キーワード抽出手法

本章では，3.2.1 項で述べた辞書に依存しない形態素と N-gram を組み合わせ，$RIDF$ を利用した重み付けによる観光キーワード抽出手法について詳細を記述する．我々が構築した提案システムは，形態素解析結果を利用するが，名詞及び複合名詞のみを観光キーワードとすると，形態素解析の誤判断や辞書登録のない新語，名詞句の抽出ができない課題がある．このため，一般的には新語や名詞句を辞書登録し対応するが，それでは人手によるコストがかかる．また，特定地域に限定せず観光情報の自動抽出・融合を行うには，人手による辞書作成は限界がある．そこで，我々は，池野ら [10] の専門用語獲得手法を参考に，形態素と N-gram を組み合わせ，$RIDF$($ResidualIDF$; 残差 IDF) を利用した重み付けによる観光キーワードの抽出手法を提案する．

ここで，観光キーワード抽出手法の詳細を述べる．3.2.1 項 (2) に示した地域サイトから観光キーワード抽出処理は，次の手順で行う．

（1）　地域サイトの文書からタグや改行除去

(注1)：http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

（2） 形態素解析器 (Mecab) による形態素解析
（3） 形態素 N-gram リストの生成 (1-gram〜5-gram)
（4） $RIDF$ による重み付け
（5） 閾値を基準に観光キーワードの抽出

処理 (1) では，地域サイトから収集した文書からタグや改行を除去し形態素解析を行う前処理を行う．

処理 (2) では，Mecab を利用した形態素解析を行い形態素を生成する．

処理 (3) では，処理 (2) で生成した形態素と出現した順序を基に，形態素 N-gram を利用した形態素 N-gram リストを生成する．ここで，$N = 5$ として形態素 N-gram リストを求めた．例として,「魚津の蜃気楼と歴史」は，形態素解析により,「魚津」・「の」・「蜃気楼」・「と」・「歴史」と分かち書きされる．この場合の形態素 N-gram リストは,「魚津」,「魚津の」,「魚津の蜃気楼」,「魚津の蜃気楼と」,「魚津の蜃気楼と歴史」,「の」,「の蜃気楼」,「の蜃気楼と」,「の蜃気楼と歴史」,「蜃気楼」,「蜃気楼と」,「蜃気楼と歴史」,「と」,「と歴史」,「歴史」の 15 通りとなる．

処理 (4) では，生成した形態素 N-gram リストに対し，$RIDF$($ResidualIDF$; 残差 IDF) を利用した重み付けを行う．これは，ポアソン分布が文書集合の一般語に対して当てはまり，キーワードに対しては当てはまらないという考え方を利用した手法である [11]．ここで，任意の形態素 N-gram を X とし，次の統計量を利用した．

Z : 収集した Web 文書数
$CF(X)$: X の Web 文書中の出現回数
$DF(X)$: X の出現する Web 文書数
$IDF(X)$: X の逆文書頻度
$\lambda(X)$: X のポアソン分布のパラメータ
$P(0;\lambda(X))$: X が 1 度も出現しない確率
$\hat{N}(X)$: X の出現頻度の推定値
$I\hat{D}F(X)$: X の IDF の推定値
$RIDF(X)$: X の $RIDF$

また，各統計量の計算式は，次のとおりである．

$$IDF(X) = \log Z - \log DF(X) \tag{1}$$
$$\lambda(X) = \frac{CF(X)}{Z} \tag{2}$$
$$P(0;\lambda(X)) = e^{-\lambda(X)} \tag{3}$$
$$\hat{N}(X) = Z(1 - P(0;\lambda(X))) \tag{4}$$
$$I\hat{D}F(X) = \log \frac{1}{1 - P(0;\lambda(X))} \tag{5}$$
$$RIDF(X) = IDF(X) - I\hat{D}F(X) \tag{6}$$

処理 (5) では，$RIDF$ の平均値と分散を利用し形態素 N-gram リストから，観光キーワードを抽出する．ここで，次の統計量を利用し，$\mu \leqq RIDF(X) < \mu + \sigma$ 内であるものを抽出し，観光キーワードとした．

μ : $RIDF$ の平均値
$\sigma^2(X)$: $RIDF(X)$ の分散
σ : $RIDF$ の標準偏差

各統計量の計算式は，次のとおりである．

$$\mu = \frac{\sum_{i=1}^{k} RIDF(X)}{k} \tag{7}$$
$$\sigma^2(X) = (RIDF(X) - \mu)^2 \tag{8}$$
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{k}(RIDF(X) - \mu)^2}{k}} \tag{9}$$
$$P(0;\lambda(X)) = e^{-\lambda(X)} \tag{10}$$

5. 提案システムの検証実験

本章では，3 章で示した提案システムの有用性を確かめるために行った検証実験について記述する．5.1 節に実験対象，5.2 節に検証実験結果を記述する．

5.1 実験対象

検証実験の実験対象地域は，本研究に関連して共同研究を行った富山県魚津市の観光情報とした [12]．また，特定地域に限定しない抽出手法としての有用性を確認するために，東京都日野市の観光情報も実験対象として，魚津市と同様の実験を行うこととした．

ここで，3.2.1 項で述べた観光情報抽出の実験を 5.1.1 項，3.2.2 項で述べた観光情報融合の実験を 5.1.2 項に示す．

5.1.1 観光情報抽出実験

3.2.1 項 (1) に示した地域サイトからの観光情報収集を行う対象は，魚津市と魚津市観光協会の公式サイトとした．収集範囲は，魚津市の Web ページ内の観光情報(注2)のトップページと魚津市観光協会公式サイト魚津たびナビ(注3)を基準として下位 2 階層を収集対象とした．また，日野市は，日野市観光協会(注4)のトップページを基準として下位 2 階層を収集対象とした．

3.2.1 項 (2) と 4 章に示した観光キーワード抽出は，地域サイトの収集対象とした魚津市の約 1,649 ページと日野市の約 1,813 ページを対象に行った．

3.2.1 項 (3) に示したブログの収集対象は，特定のジャンルに特化せず多くの層の様々な話題を収集できる Yahoo Japan! が運営する Yahoo!ブログ(注5)を対象とすることとした．ブログの観光情報抽出対象は，魚津市は，Yahoo!ブログの検索においてキーワード「魚津」で検索されたブログ記事の新着 500 ページとした．同様に，日野市は，Yahoo!ブログの検索においてキーワード「日野」で検索されたブログ記事の新着 500 ページとした．

5.1.2 観光情報融合実験

3.2.2 項 (1) に示した地域サイトの分類は，収集した地域サイトを表 1 に示した「見る・遊ぶ」,「イベント・祭り」,「食べる・泊まる」,「お土産・特産品」,「自然・文化」の 5 つのカテゴリに事前定義語を利用し分類した．

3.2.2 項 (2) に示したブログの関連性判断は，実験対象の地

(注2) : http://www.city.uozu.toyama.jp/topVisit.aspx
(注3) : http://www.uozu-kanko.jp/
(注4) : http://shinsenhino.com/
(注5) : http://blogs.yahoo.co.jp/

域サイトの情報から抽出した観光キーワードを基準に，収集したブログページ 500 ページ中の 50 ページを学習データとして線形判別関数を生成した．生成した線形判別関数を利用し，検索対象地域である魚津市と日野市のブログが観光情報と関連性があるかを判断する実験を行った．

3.2.2 項 (3) に示したブログの分類は，3.2.2 項 (2) の関連性判断により観光情報に関連があると判断したブログページを表 1 の 5 つのカテゴリに分類する．このとき，3.2.2 項 (1) の地域サイトの分類により 5 つのカテゴリに分類された地域サイトページ群の観光キーワードを基準に分類を行った．

5.2 実験結果

5.1 節で示した実験対象地域である富山県魚津市と東京都日野市の実験結果を示す．5.1.1 項の観光情報抽出実験の結果を項，5.1.2 項の観光情報融合実験の結果を項に示す．

5.2.1 観光情報抽出実験結果

(1) 地域サイトからの観光情報収集は，地域対象の魚津市と魚津市観光協会の地域サイトを基準に下位 2 階層の 1,649 ページを対象として収集を行った．収集先が PDF や画像の場合は除外し，実際に収集したページは 1,299 ページとなった．同様に日野市は，日野市観光協会のサイトを基準に下位 2 階層の 1,813 ページを対象として収集を行い，PDF や画像のページを除外し 1,731 ページとなった．この実験により，収集対象の基準としたページから下位 2 階層を収集すると地域サイトによってページ数は異なるがサイトの構成から約 1,500 ページ前後の Web 文書集合が得られることを確認した．

(2) 観光キーワード抽出は，(1) により収集した地域サイト (魚津市 : 1,299 ページ，日野市 : 1,731 ページ) の Web 文書集合から $N = 5$ となる形態素 N-gram を抽出し，$RIDF$ による重み付けを行った．魚津市・日野市から抽出した形態素 N-gram は，それぞれの地域で約 60 万語前後となった．この約 60 万語の形態素 N-gram の $RIDF$ の平均値と標準偏差から，$\mu \leqq RIDF(X) < \mu + \sigma$ となる形態素 N-gram を魚津市・日野市それぞれから約 30 万語を抽出した．さらに，形態素 N-gram から記号で始まるものを除去した約 15,000 語を魚津市・日野市の各地域ごとにそれぞれ抽出し，これを観光キーワードとして利用することとした．本手法において抽出した約 15,000 語の内，実際に観光キーワードになりうると考えられる語は，人手により約 40%の約 6,000 語であることを確認した．この実験により，約 1,500 ページの文書集合から形態素 N-gram と $RIDF$ の重み付けにより約 15,000 語の観光キーワードを取得できることが確認できた．また，約 40%の語が観光キーワードとして有用であることも併せて確認した．

ここで，表 2 に，魚津市と日野市の観光キーワードとして取得したそれぞれ約 15,000 語の中で，観光キーワードとして利用可能であると人手で判断した $RIDF$ の高い上位 20 件を示す．魚津市では，寿司・鮮魚・りんごなどの名物や代表的な交通手段となる JR 魚津駅・市民バス・タクシー・車などが上位に抽出された．また，蜃気楼など特に有名な観光資源についても 20 位には入らないものの上位に抽出されている．また，辞書には登録されていない魚津市のゆるキャラとして作られた「ミラたん」やイベントである「まるまる魚津」などの観光キーワードの抽出も確認できた．これに対し日野市では，最も有名な観光資源となる新撰組に関するイベントや人名がキーワードの上位に位置している．また，移動手段となる多摩モノレールや京王線の駅名，市内連絡バスなどのキーワードが抽出されている．さらに，日本野鳥の会やブルーベリーといったキーワードも抽出されている．この実験結果により，各地域の観光キーワードとして必要となる語が形態素 N-gram に含まれ，$RIDF$ による重み付けによる特徴語を観光キーワードとして抽出できることを確認した．また，魚津市と日野市それぞれでの実験により地域特有の語を抽出できている点から，本手法により地域に偏らない観光キーワードの抽出ができることを確認できた．

表 2 $RIDF$ による観光キーワード上位 20 件
Table 2 Sightseeing keyword high rank 20 cases by the $RIDF$ level.

魚津市	日野市
JR 魚津駅より徒歩	系統バス
JR 魚津	アクセス多摩モノレール甲州街道
ビジネス	パレード新撰組
JR 魚津駅	日野宿会場
車で	宇宙
ファイル	市内連絡バス
市民バス	原田左之
寿司	ひの新撰組まつり
人数	三番隊
散策ガイド	十番隊隊長
散策ガイドマップ	八番隊
散策	高幡会場
イベント内容	日本野鳥の会
お話の会	源氏物語
海の駅・蜃気楼	モノレール甲州街道駅より
タクシー	甲州街道駅より徒歩
レンタサイクル”みらくる	平山季重
地方発送	ブルーベリー
鮮魚	高幡不動駅より
りんご	藤堂平助

(3) ブログからの情報収集は，Yahoo!ブログからキーワード「魚津」と「日野」のそれぞれのキーワードで検索されたブログ記事の新着 500 ページを収集したことを確認した．現時点では実験のため収集件数を 500 ページとしているが，今後は収集件数を増やすことや別のブログサイトを収集対象とするなど，より有用な観光情報を多く集めることを検討している．

5.2.2 観光情報融合実験結果

(1) 地域サイトの分類は，3.2.2 項の表 1 に示す 5 つのカテゴリに収集した地域サイトの各ページを分類した．表 3 に，魚津市の地域サイトの分類結果を示す．各カテゴリに分類した地域サイトの適合率の平均は 76.8%となり，地域サイトとしては情報量が少ない「食べる・泊まる」のカテゴリを除き誤りが少ない結果となった．しかし，再現率の平均は 27.5%となり，事前定義語が少ないことによる取りこぼしが多い結果となった．

表 4 に，日野市の地域サイトの分類結果を示す．魚津市の実

表 3　地域サイトの分類結果 (魚津市)
Table 3　Classification of web pages in the regional sites(Uozu-City).

カテゴリ	見る 遊ぶ	イベント 祭り	食べる 泊まる	お土産 特産品	自然 文化
地域サイト (適合率)	88.9%	76.0%	85.0%	51.9%	89.4%
地域サイト (再現率)	5.2%	29.7%	26.2%	30.4%	46.1%

表 4　地域サイトの分類結果 (日野市)
Table 4　Classification of web pages in the regional sites(Hino-City).

カテゴリ	見る 遊ぶ	イベント 祭り	食べる 泊まる	お土産 特産品	自然 文化
地域サイト (適合率)	83.3%	86.4%	66.6%	87.5%	100.0%
地域サイト (再現率)	15.6%	29.7%	1.6%	10.9%	15.6%

験結果と同様に適合率の平均は，81.8%と，日野市の観光情報として極めて少ない「食べる・泊まる」のカテゴリ以外は誤りが少ない結果となった．しかし，再現率の平均は 22.2%となり，魚津市同様に取りこぼしが多い結果となった．

よって，地域サイトの分類の実験結果から，各地域サイトの分類は，適合率が高く，再現率が低くなる結果が得られた．これは，現在の地域サイトの分類手法が，表 1 に示す事前定義語を基準に行っているためである．このため，事前定義語に該当した地域サイトは，分類が正しく行われる傾向がある．しかし，事前定義語に該当しない場合は，カテゴリに分類されず該当カテゴリがないページとなるため再現率が低くなる．そこで，事前定義語を各地域によって適切に設定し該当カテゴリがないページを減らし再現率を向上させるには，事前定義語自体の自動生成を行うなど人手を介さない手法を今後検討する必要がある．

(2) ブログの関連性判断の実験は，ブログの収集対象とした各地域のブログ 500 ページから学習データとして 50 ページ，実験データとして 100 ページを利用し行った．この実験において収集したブログの中で実際に観光情報に関連があるページは，人手による確認で魚津市が約 50%に対し，日野市は約 30%と非常に少ない件数であった．表 5 に，魚津市と日野市の線形判別結果を示す．適合率は，魚津市と日野市それぞれで，50.0%・59.5%となった．また，再現率は，魚津市と日野市でそれぞれ，100%，86.7%となった．よって，本手法による線形判別式は，適合率が低く再現率が高い判断を行う結果となったため，収集したブログに観光情報に関連したページ数が少ない場合でも取りこぼしを減らすことができると考えられる．しかし，観光情報が非常に多い地域に対して本手法の線形判別が適応可能か検証する必要がある課題を残している．

(3) ブログの分類実験は，(2) の実験により観光情報に関連性があると判断したブログページを 5 つのカテゴリへ分類し検証

表 5　形態素 N-gram による線形判別結果
Table 5　Linear discriminant analysis result by morpheme N-gram.

	適合率	再現率
魚津市	50.0%	100.0%
日野市	59.5%	86.7%

表 6　形態素 N-gram によるブログの分類結果 (魚津市
Table 6　Classification of web pages in the blog sites by morpheme N-gram(Uozu-city).

カテゴリ	見る 遊ぶ	イベント 祭り	食べる 泊まる	お土産 特産品	自然 文化
ブログ (適合率)	60.0%	13.0%	30.0%	10.5%	43.5%
ブログ (再現率)	36.0%	20.0%	35.3%	100.0%	47.6%

した．表 6 に魚津市に関連性があると判断したブログ 50 ページの分類結果を示す．ここで分類の基準は，表 3 に示した地域サイトの分類結果を利用し，各カテゴリに分類された地域サイトページ群の観光キーワードとした．本手法による魚津市に関連したブログの適合率と再現率の平均は，31.4%，47.9%となった．よって，誤認識した地域サイトページに含まれる観光キーワードも含めて分類の基準とするため，適合率は地域サイトの分類と比較し低下する．しかし，他のカテゴリと比べ極端に情報が少ないカテゴリである「お土産・特産品」に関するブログページを分類できていることから，地域サイトの分類と比較すると取りこぼしが少ないと考えられる．

また，表 7 に日野市に関連性があると判断したブログ 50 ページの分類結果を示す．分類の基準は，表 4 に示した地域サイトの分類結果を利用し，各カテゴリに分類された地域サイトページ群の観光キーワードとした．ここで，本実験において収集した日野に関連するブログには,「食べる・泊まる」・「お土産・特産品」の情報を得ることができなかったため，2 つのカテゴリについてはデータ無しとした．本手法による日野市に関連したブログの適合率と再現率の平均は，55.9%，31.2%となった．この結果から，分類された地域サイトページ群の観光キーワードを基準とするため，魚津市と同様に，適合率は地域サイトの分類と比較し低下する．しかし，再現率は地域サイトの分類と比較し高くなることから，地域サイトページ群に関連するブログを取りこぼしが少なく分類できていると考えられる．

この実験結果から，本手法により，特定の地域によらずブログの件数が少ない場合でも地域サイトの分類結果を利用しブログの分類を行うことが可能である．今後，分類方法についてはさらに議論が必要であるが，形態素 N-gram を利用した観光キーワードをブログの分類に利用できることを確認した．

6. 考察と今後

本稿において，形態素 N-gram を利用し $RIDF$ による重み付けを利用した観光キーワード抽出手法を示し，魚津市と日野市の観光情報に対して適用した実験結果を示した．この実験結

表 7　形態素 N-gram によるブログの分類結果 (日野市)

Table 7　Classification of web pages in the blog sites by morpheme N-gram(Hino-city).

カテゴリ	見る 遊ぶ	イベント 祭り	食べる 泊まる	お土産 特産品	自然 文化
ブログ (適合率)	86.4%	25.0%	–	–	56.1%
ブログ (再現率)	42.2%	21.4%	–	–	30.0%

果により，形態素 N-gram を利用し $RIDF$ による重み付けを行う観光キーワードの抽出手法は，特定の地域によらず地域の観光情報の特徴を抽出することができることを確認した．また，自動抽出した観光キーワードを利用した地域サイトの分類，ブログの関連性判断，ブログの分類は，改善点は多いものの特定の地域によらず観光情報を正しく判断し分類可能であることを確認した．

このため，本稿で提案した形態素 N-gram による観光キーワードの抽出手法は，辞書登録などの人手を必要とせず低コストで特定の地域に偏らず利用可能な観光情報の自動抽出・融合手法となることを確認した．関連性判断や分類の手法についてはさらなる議論を必要とするが，Web 上の大量の情報から特定地域の有用となる観光情報の自動抽出・融合を行うことができると考えられる．今後は，提案システムを更に改善し，特定地域の観光情報の自動抽出・融合の精度向上と Web 上の大量の情報から有用な情報の抽出・融合を行うことができる汎用的な手法を検討していきたい．

文　　献

[1] 遠藤 雅樹, 大野 成義, 石川 博 : 地域サイトからの観光キーワードの自動抽出と関連情報の融合，第 158 回データベースシステム研究発表会, A1-2, 2013.

[2] 斎藤 一 : Web における観光情報の提供と分析，人工知能学会誌, 26 巻 3 号, pp.234-239, 2011.

[3] 三田村 保, 岩佐 渉, 湯川 恵子, 大堀 隆文 : ブログを利用した観光情報の調査分析，観光情報学会論文誌, Vol.4,No.1, pp.57-65, 2008.

[4] 守屋 豊, 井出 明 : テキストマイニングを用いた観光の言説分析，情処学研報, 2008-DD-64,No.8, pp.55-60, 2008.

[5] 石野 亜耶, 難波 英嗣, 竹澤 寿幸 : 旅行ブログエントリからの観光情報の自動抽出，日本知能情報ファジィ学会誌, Vol.22, No.6, pp.667-679, 2010.

[6] 寺西 拓也, 野村 達二, 平山 智子, 石野 亜耶, 難波 英嗣, 竹澤 寿幸 : 観光ガイドブックへの旅行ブログエントリと質問応答コンテンツの対応付け，言語処理学会, 第 18 回年次大会発表論文集, pp.333-336, 2012.

[7] 森本 泰貴, 藤本 典幸, 長屋 務, 出原 博, 萩原 兼一 : Web を対象としたロボット型住所関連情報検索システムの開発，信学論, Vol.J90-D, No.2, pp.245-256, 2007.

[8] 遠藤 雅樹, 大野 成義 : 地域サイト及びブログからの観光情報の自動抽出と融合，DEIM Forum 2013, F9-2, 2013.

[9] 小作 浩美, 内山 将夫, 井佐原 均, 河野 恭之, 木戸出 正継 : 新聞記事コーパスでの単語出現特徴を利用した観光イベント情報の検索支援，人工知能学会論文誌, Vol.19, No.4D, pp.225-233, 2004.

[10] 池野 篤司, 濱口 佳孝, 山本 英子, 井佐原 均 : Web 文書集合からの専門用語獲得，情報処理学会論文誌, Vol.47, No.6, pp.1717-1727, 2006.

[11] Church，K. W. and Gale，W. A. :Poisson mixtures，Journal of Natural Language Engineering, Vol.1, No.2, pp.163-190, 1995.

[12] 山中 光定, 高尾 和志, 遠藤 雅樹 : 共同研究「市民バスロケーションシステムの開発」，第 20 回職業能力開発研究発表講演会, 3-15, 2012.