Vol.2013-DBS-158 No.2
2013/11/26

地域サイト及びブログの観光情報融合のための キーワード自動抽出手法の検討

遠藤 雅樹[1,2,a)]　大野 成義[1,b)]　石川 博[2,c)]

概要：近年，Web 上に存在する大量の情報から有用な情報を取り出す技術が注目されている．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有用な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．ここでは，Web 上の地域サイト情報と地域サイトに関連するブログ情報から観光情報を自動抽出・融合する際に必要となる，各地域における観光スポットやイベントなどの観光キーワードの自動取得について議論する．一般的に，辞書登録のない未知語を含む観光キーワードは，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．そこで，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する．本手法により，特定地域に限定せず低コストに検索地域の有用な観光情報を取り出せる可能がある．本稿では，地域サイトの情報から未知語を含む観光スポットなどの観光キーワードを自動取得する手法の提案とその有用性について報告する．

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，様々な種類の大量の情報が Web 上に置かれる傾向にある．そして，Web 上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 に置かれることが予想されている．このため，Web の情報量の増加に伴い，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しても Web 上には大量の情報が点在している．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いる．そのため，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．

[1] 職業能力開発総合大学校
Polytechnic University
[2] 首都大学東京
Tokyo Metropolitan University
[a)] endou@uitec.ac.jp
[b)] ohno@uitec.ac.jp
[c)] ishikawa-hiroshi@tmu.ac.jp

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難しくなっている．これは，検索エンジンなどから得られた検索結果から参考になる観光情報を抽出し，関連した観光情報同士を融合する処理は，システム化されていないためである．ここで，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判断し，どの観光スポットについて記述されているのか内容を確認することを指す．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情報同士を結び付けて観光の参考となる情報にまとめる作業を指す．この情報抽出と情報融合は，システム化されておらず人手で行う作業である．そのため，Web 利用者個々の情報処理能力に依存している．よって，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．仮に検索結果により抽出された情報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．さらに，Web 上の情報量が増加することで益々困難となる．

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・融合する手法を検討している．特に，人手による辞書作成など頻繁なメンテナンスを必要と

しない低コストに運用可能で特定地域に限定せず観光情報を自動抽出・融合可能な手法を議論している．

本稿では，Web 上の検索対象とした地域サイトに存在する情報から観光スポットやイベントなどの観光キーワードを自動取得する手法を提案する．また，実験により，辞書を利用せず特定地域の観光キーワードを取得できる提案手法の有用性について報告する．

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 [1].

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 [2] 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 [3] 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 [4][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や旅行ガイドブックの情報拡張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．しかし，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，低コストに最新の観光情報を収集可能である．また，人手によるコストをかけず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辞書の充実により精度を向上させる手法とは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有用な情報を取り出す技術につながる可能性がある．

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 [8] 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ロコミ情報なども抽出し関連する情報同士をまとめる情報融合を行う点に特徴がある．これは，観光スポットやイベントなどの情報に対して付加価値のあるロコミ情報などを関連付けることができるため，抽出した情報を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，3.3 節に提案システムの性能を記述する．

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が利用することを想定した観光情報の自動抽出・融合を行うシステムを提案する．

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，Web を利用した情報収集は，次の手順で検索を行うことが予想される．

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．

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．しかし，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，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

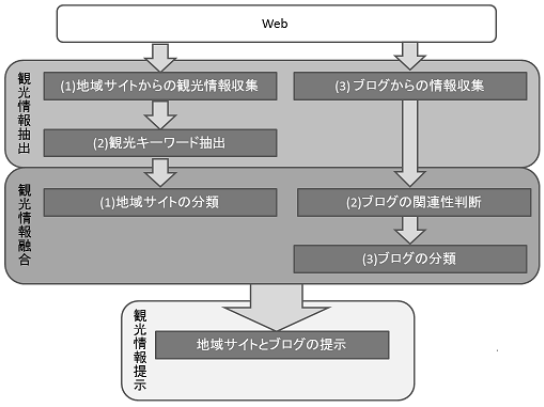

図 1 システム処理手順
Fig. 1 System processing procedure.

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 [9]．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 節に観光情報抽出，3.2.2 節に観光情報融合，3.2.3 節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，次の処理を行う．

(1) 地域サイトからの観光情報収集
(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，検索対象とした観光地の自治体や観光協会の提供する地域サイトから観光情報を収集する．収集する情報は，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情報である．一般的に地域サイトの情報は信頼性が高いため，地域の観光名所や施設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特徴語の抽出を行う．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．観光キーワードとなる特徴量は，形態素解析器 (Mecab[*1]) による形態素解析結果から複合名詞を取り出し，$TF\text{-}IDF$ を基準とした重み付けを行った．また，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 節の観光情報融合を行う．

表 1 カテゴリ含有語
Table 1 Category component word.

カテゴリ	カテゴリ含有語		
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

(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．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含有語を定義した．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．各カテゴリ含有語は，合計約 60 語を定義した．表 1 に，各カテゴリとカテゴリ含有語の例を示す．

観光情報融合は，3.2.1 節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事前定義した 5 つのカテゴリのどのカテゴリに属するかをカテゴリ含有語の出現頻度を基準に分類する．

(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．関連性判断は，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 件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 (3) ブログの分類に利用する．

(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，5 つのカテゴリに分類された地域サイトのページに

[*1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

含まれる観光キーワードを利用し，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．この処理により，観光情報抽出により取得した地域サイトとブログの内容が関連した情報同士を結びつける．

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．

3.3 提案システムの検証実験

我々は，提案システムのプロトタイプを製作し観光情報の抽出・融合手法の有効性を検証する実験を行った．3.3.1 節に実験対象の概要，3.3.2 節に検証実験の結果，3.3.3 に提案システムの問題点を記述する．

3.3.1 実験対象の概要

検索対象地域は，本研究に関連して共同研究を行った富山県魚津市とした [10]．地域サイトの収集対象は，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[*2]のトップページを基準にその下位ページと魚津市観光協会公式サイト魚津たびナビ[*3]内の合計約 100 ページを対象とした．ここで，各サイトから外部ドメインへのリンクは抽出対象外としている．

また，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[*4]を対象とすることとした．ブログの観光情報抽出対象は，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 件とした．

3.3.2 検証実験の結果

本節では，3.3.1 節の実験対象に対して行った検証実験結果を示す．3.3.2.1 に 3.2.1 節の観光情報抽出，3.3.2.2 に 3.2.2 節の観光情報融合についての実験結果をそれぞれ記述する．

3.3.2.1 観光情報抽出の実験結果

(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から約 100 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 78 件となった．

(2) 観光キーワード抽出は，収集した地域サイト 78 件から行った．収集した地域サイトページの形態素解析結果から複合名詞を取り出し，*TF-IDF* を利用した重み付けを行い観光キーワードとした．表 2 に *TF-IDF* 値の高い上位 20 件を示す．抽出した観光キーワードは約 1 万語である．

[*2] http://www.city.uozu.toyama.jp/topVisit.aspx
[*3] http://www.uozu-kanko.jp/
[*4] http://blogs.yahoo.co.jp/

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

観光キーワード	*TF-IDF*
海	0.704
定番商品	0.627
工芸品	0.627
イチオシ	0.627
オリジナル	0.627
海産物	0.624
魚津漆器	0.616
伝統	0.613
漆器	0.612
特産品	0.566
とき	0.557
新鮮	0.556
幸	0.546
歴史	0.545
今	0.537
魚津	0.519
map	0.507
地元	0.482

表 3 地域サイトの分類結果
Table 3 Classification of web pages in the regional sites.

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%

その中で，実際に観光キーワードになりうると考えられる語は，約 30%の約 3,000 語であることを確認している．尚，実験結果の検証は，魚津市の観光情報を知らない 3 名と魚津市の観光情報について知っている 1 名の 4 名の人手により判断した．以降に示す検証についても同様である．

(3) ブログからの情報収集は，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 件を収集した．

3.3.2.2 観光情報融合の実験結果

(1) 地域サイトの分類は，3.3.2 節の表 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，地域サイトページ 78 件の観光情報の分類結果を示す．各カテゴリに分類した地域サイトの適合率と再現率の平均は，それぞれ 78.2%，27.5%となった．事前定義語を利用する地域サイトの分類手法は，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予定である．

(2) ブログの関連性判断は，ブログの収集対象とした 500 ページから学習データとして 50 件，実験データとして 56 件を利用し行った．表 4 に，実験データ 56 件の線形判別結

表 4　線形判別結果
Table 4　Linear discriminant analysis result.

人手＼機械	関連性有	関連性無	正判別率
関連性有	5	6	45.5 %
関連性無	5	40	88.9 %
		全体の正判別率	80.4 %

表 5　ブログの分類結果
Table 5　Classification of web pages in the blog sites.

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)	71.4%	21.4%	0.0%	0.0%	52.5%
ブログ (再現率)	12.5%	11.1%	0.0%	0.0%	75.0%

果を示す．実験データ全体の正判別率は，80.4%となった．

(3) ブログの分類は，ブログの関連性判断により関連性有としたブログページを 5 つのカテゴリに分類した．分類には表 4 に示す 5 つのカテゴリに分類された地域サイトページに含まれる観光キーワードを基準に行った．表 5 にブログ 500 件の観光情報の分類結果を示す．各カテゴリに分類したブログの適合率と再現率の平均は，それぞれ 29.1%，19.7%となった．

3.3.3　提案システムの問題点

3.3.2 節にプロトタイプとして構築した提案システムの検証実験の結果を示した．提案システムにより観光情報の自動抽出・融合を行うシステムとして有用となる可能性は確認できた．しかし，提案システムの適合率再現率の改善に，観光情報抽出処理で行う観光キーワードの抽出精度の向上が必要となると考えられる．提案システムでは，観光キーワード抽出には，形態素解析結果から複合名詞を取り出し，$TF\text{-}IDF$ による重み付けを行い抽出している．しかし，形態素解析の誤判断や新語，名詞句の抽出ができない課題がある．このため，新語や名詞句を辞書へ登録し対応する手法が一般的であるが，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には限界がある．そこで，我々は，池野ら [11] 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抽出処理の改善手法を提案する．次章に，この改善手法と改善手法の適用による提案システムの性能向上について記述する．

4.　観光キーワード抽出処理の改善

本章では，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の改善手法を提案する．我々が 3.2.1 節で述べた提案システムの観光キーワード抽出手法は，複合名詞を取り出し $TF\text{-}IDF$ による重み付けを行い抽出する．しかし，3.3.3 節で述べたとおり，形態素解析の誤判断や新語，名詞句の抽出ができない課題がある．ここでは，観光キーワード抽出を行う改善手法を提案し，4.1 節に観光キーワード抽出の改善手法，4.2 節に改善手法の検証実験について示す．

4.1　観光キーワード抽出の改善手法

本節では，3.2.1 節で述べた観光キーワード抽出の改善手法について示す．ここで，地域サイトから収集した文書を基に観光キーワードの抽出を次の手順で行う．

(1) 地域サイトの文書からタグや改行除去
(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

3.2.1 節の提案システムにおいて複合名詞と $TF\text{-}IDF$ による重み付けを利用していたが，改善手法では上記の (3) から (5) の手法に変更した．

(3) では，(2) で収集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 リストを生成した．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．

(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 [12]．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
$\hat{IDF}(X)$：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．

$$IDF(X) = \log Z - \log DF(X) \quad (1)$$

$$\lambda(X) = \frac{CF(X)}{Z} \quad (2)$$

$$P(0;\lambda(X)) = e^{-\lambda(X)} \quad (3)$$

$$\hat{N}(X) = Z(1 - P(0;\lambda(X))) \quad (4)$$

$$I\hat{D}F(X) = \log \frac{1}{1 - P(0;\lambda(X))} \quad (5)$$

$$RIDF(X) = IDF(X) - I\hat{D}F(X) \quad (6)$$

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，$RIDF(X)$ が μ 以上で $RIDF(X)$ が $\mu \pm \sigma$ 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．

$$\mu = \frac{\sum_{i=1}^{k} RIDF(X)}{k} \quad (7)$$

$$\sigma^2(X) = (RIDF(X) - \mu)^2 \quad (8)$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{k}(RIDF(X) - \mu)^2}{k}} \quad (9)$$

$$P(0;\lambda(X)) = e^{-\lambda(X)} \quad (10)$$

4.2 改善手法の検証実験

本節では，4.1 で示した改善手法の有効性を検証する実験について記述する．4.2.1 節に改善手法の実験対象，4.2.2 節に改善手法の検証実験結果，4.2.3 節に提案システムへの適用結果を記述する．

4.2.1 改善手法の実験対象

改善手法の検証実験は，3.3.1 節に記述した実験対象と同様である．よって，検索対象地域は，魚津市の Web ページ内の観光情報のトップページを基準にその下位ページと魚津市観光協会公式サイト魚津たびナビ内の合計約 100 ページが対象であり，各サイトから外部ドメインへのリンクは抽出対象外としている．

また，ブログの収集対象は，Yahoo Japan!が運営する Yahoo!ブログを対象に，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 件である．

改善手法での変更点は，収集した地域サイトページから観光キーワードを抽出する際に，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い抽出を行った点である．次節に TF- による重み付けでなく 4.1 節の (3) から (5) の処理を行った改善手法の検証実験結果について記述する．

4.2.2 改善手法の検証実験の結果

本節では，4.2.1 節の実験対象に対して行った改善手法による観光キーワード抽出の検証実験結果を示す．4.2.2.1 に 4.1 節の (3) から (5) の観光キーワード抽出についての実験結果，4.2.2.2 に観光キーワード抽出例を記述する．

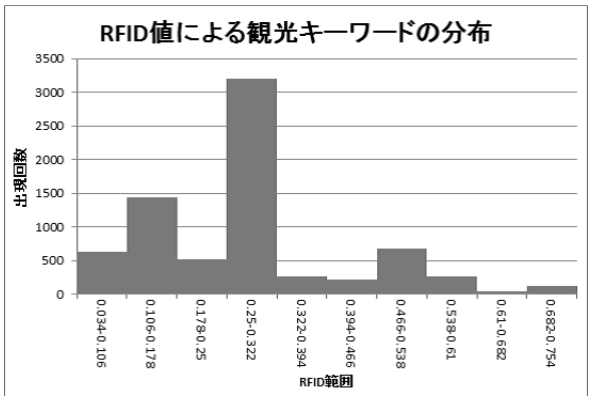

図 2 $RIDF$ 値による観光キーワードの分布
Fig. 2 Distribution of tourism keywords by RIDF value.

4.2.2.1 改善手法による観光キーワード抽出の実験結果

4.1 節の (3) から (5) についての実験結果を示す．

(3) 形態素 N-gram リストの生成は，地域サイトより収集した約 100 ページの Web 文書集合から形態素解析により取得した形態素を利用した．生成された 1-gram から 5-gram の形態素 N-gram リストは，約 11 万語となった．

(4)$RIDF$ による重み付けは，約 11 万語の形態素 N-gram リストに対して行った．形態素 N-gram リストの $RIDF$ の平均値と標準偏差はそれぞれ，$\mu = 0.032$，$\sigma = 0.224$ となった．

(5) 閾値を基準に観光キーワードの抽出は，形態素 N-gram リストの絞込みを行った．その際，$RIDF(X) \geq \mu$，かつ，$\mu - \sigma < RIDF(X) < \mu + \sigma$ を基準とした．その結果，形態素 N-gram リストから約 1 万語を抽出した．さらに，観光キーワードとして利用するために，抽出した形態素 N-gram リストから記号で始まるものを除去し，約 7,400 語を観光キーワードとして利用することとした．図 2 に，抽出した約 7,400 語の形態素 N-gram の出現回数ヒストグラムを示す．改善手法において抽出した約 7,400 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 2,800 語であることを確認した．

4.2.2.2 改善手法による観光キーワード抽出例

4.2.2.1 の実験結果から抽出された観光キーワード例を表 6 に示す．表 6 は，観光キーワードとして利用可能な $RIDF$ の高い上位 20 件である．

ここで，3.3.2 節の表 2 の実験結果である複合名詞と TF-IDF による重み付け手法と比較した．その結果，3.3.2 の手法では抽出できない「JR 魚津駅より徒歩」，「海の駅・蜃気楼」のような名詞句を取得できていることが確認できた．

また，単純な名詞連結により構成した複合名詞と TF-IDF による重み付けで抽出できた重要な観光キーワードは，$RIDF$ を利用した提案手法においても網羅できていることを人手により確認した．加えて，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

表 6 *RIDF* による観光キーワード上位 20 件
Table 6 Sightseeing keyword high rank 20 cases by the *RIDF* level.

観光キーワード	*RIDF*
JR 魚津駅より徒歩	0.750
JR 魚津	0.749
ビジネス	0.743
JR 魚津駅	0.741
車で	0.734
ファイル	0.733
市民バス	0.729
寿司	0.720
人数	0.710
散策ガイド	0.707
散策ガイドマップ	0.707
散策	0.691
イベント内容	0.685
お話の会	0.685
海の駅・蜃気楼	0.685
タクシー	0.685
レンタサイクル”みらくる	0.685
地方発送	0.685

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．

実験結果から，形態素 N-gram と *RIDF* を利用した観光キーワード抽出の改善手法は，複合名詞と *TF-IDF* による重み付けを利用した比較手法に比べ観光キーワードとして利用可能な語をより正確に抽出できることを確認した．さらに，改善手法によって，複合名詞だけでなく名詞句などの観光キーワードも抽出可能であり，比較手法に比べノイズが少なく観光キーワードとして有用な語を取得することに成功した．よって，図 1 に示す提案システムの観光キーワード抽出処理に改善手法を適用することで，観光情報の自動抽出・融合を向上させることができると考えられる．

次節で提案システムへの改善手法の適用結果を示す．

4.2.3 提案システムへの適用結果

本節では，3 章で述べた提案システムに 4.1 節の形態素 N-gram に対して *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適用した実験結果を示す．4.2.3.1 に実験内容，4.2.3.2 に観光情報融合の実験結果を示す．

4.2.3.1 実験内容

実験内容は，観光キーワードの抽出手法をのみを変更し，3.3 節と同様として，ブログの関連性判断とブログのカテゴリへの分類の実験を行った．4.2.3.2 に実験結果を示す．

4.2.3.2 観光情報融合の実験結果

3.2.2 節に示した観光情報融合の (1) から (3) の改善手法を利用した場合の実験結果を示す．

(1) 地域サイトの分類は，3.3.2 節に示した分類手法と変更がないため，表 3 と同様である．この点については，事前定義語の自動生成など人手を介さない手法について今後の検討項目としている．

表 7 形態素 N-gram による線形判別結果
Table 7 Linear discriminant analysis result by morpheme N-gram.

人手＼機械	関連性有	関連性無	正判別率
関連性有	17	0	100.0 %
関連性無	17	22	43.5 %
		平均	60.7 %

(2) ブログの関連性判断は，ブログの収集対象とした 500 ページから学習データとして 50 件，実験データとして 56 件を利用し行った．表 7 に，改善手法での実験データ 56 件の線形判別結果を示す．学習データを利用し生成した線形判別式の正判別率は，74.0%となった．この線形判別式を実験データに適用した結果，実験データ全体の正判別率は 60.7%となった．これは，3.3.2 節の表 4 の複合名詞と *TF-IDF* による重み付けを利用した場合の 80.4%に対し劣っており，精度が下がる結果となった．しかし，提案手法は，人手によって関連性があると判断したデータを正しく判断する正判別率が 100.0%となり，比較手法の正判別率 45.5%に対して大きく改善し，再現率が向上したこととなる．これは，観光情報を見落とさないという意味で重要である．

この結果から，形態素 N-gram を利用した改善手法は，複合名詞を利用する場合と比べ，より確実に関連性判断ができると考えられる．関連性判断手法として線形判別以外の手法を検討するなどの議論の余地はあるが，辞書作成など人手をかけず形態素 N-gram を利用して関連性判断が可能となることを確認した．

(3) ブログの分類は，改善手法により分類された地域サイトの観光キーワードを基準に行った．表 8 にブログ 500 件の観光情報のカテゴライズ結果を示す．各カテゴリに分類されたブログの適合率と再現率の平均は，31.4%，47.9%となった．改善手法は，3.3.2 節の表 5 に示した複合名詞と *TF-IDF* による重み付けを利用する場合には抽出できなかった該当ブログ件数が少ないカテゴリ「食べる・泊まる」，「お土産・特産品」のブログを抽出することに成功した．これは，形態素 N-gram を利用し取得した観光キーワードに，複合名詞では取得できなかった「魚津のビジネスホテル」や「白エビのから揚げ」などが抽出されたためブログを分類する際の適切な語となったためであると考えられる．よって，ブログのカテゴリへの分類方法についてはさらに議論が必要であるが，改善手法による形態素 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．

表 8 形態素 N-gram によるブログのブログ結果
Table 8 Classification of web pages in the blog sites by morpheme N-gram.

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%

5. 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の改善手法を提案した．また，提案した改善手法を観光情報の自動抽出・融合を行うシステムに適用した際の実験結果も示した．この実験により，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，複合名詞と TF-IDF を利用する手法と比べノイズを約 3 割抑えることができた．また，自動抽出した観光キーワードは，抽出語の約 4 割が観光キーワードとして有効であることを確認し，複合名詞を利用した手法に比べ正確な観光キーワードを自動抽出できたことを確認した．さらに，改善手法により抽出した観光キーワードを利用したブログの関連性判断及び分類の実験では，複合名詞を用いた手法に比べ特定地域の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．

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汎用性のある観光情報の自動抽出・融合に適用できる手法であることを確認した．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．

参考文献

[1] 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.
[2] 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.
[3] 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.
[4] 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.
[5] 石野 亜耶, 藤井 一輝, 藤原 泰士, 前田 剛, 難波 英嗣, 竹澤 寿幸旅行ブログエントリと質問応答コンテンツを利用した旅行ガイドブックの情報拡張，WebDBForum2012, B2-3, 2012.
[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] 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.
[9] 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.
[10] 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).
[11] 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.
[12] Church，K. W. and Gale, W. A. : Poisson mixtures, Journal of Natural Language Engineering, Vol.1, No.2, pp.163-190,1995.