地域限定性を考慮した情報推薦における語句抽出の傾向分析とノイズの除去

西崎 剛司† 奥 健太† 服部 文夫†

†立命館大学情報理工学部

〒525-8577 滋賀県草津市野路東１丁目１番１号

E-mail: †{cc008077@ed, oku@fc, fhattori@is}.ritsumei.ac.jp

あらまし 現地でしか利用できないスポットを推薦する地域限定性を考慮した情報推薦方式の開発に取り組んでいる．この推薦方式の実現に向けて，地域限定性の高い語句を抽出する方式を提案し，その有用性を確認した．しかしながら，抽出される語句には推薦において有用でないようなノイズとなる語句も多く含まれている．本研究では，提案方式により抽出された語句の傾向を分析したうえで，そのノイズを除去する方法について検討する．

キーワード

Takashi NISHIZAKI† Kenta OKU† and Fumio HATTORI†

†Information Science & Engineering, Ritsumeikan University

1-1-1 Nojihigashi, Kusatsu-shi, Shiga, 525-8577 Japan

E-mail: †{cc008077@ed, oku@fc, fhattori@is}.ritsumei.ac.jp

1. はじめに

特定のエリアの中から条件に合致したスポットを検索するローカルサーチ（地域情報検索）がある．ローカルサーチのサービスとして，Google マップ[1]やYahoo!地図[2]などが一般公開されている．利用者が，住所や駅名などの場所および「居酒屋」や「ランチ」などのキーワードを条件として指定することにより，その場所から距離が近く，かつキーワードに合致するスポットが地図上に表示される．

しかし，我々は距離の近さやキーワードに基づく検索が必ずしも利用者にとって有用であるとは限らないと考え，地域限定性を考慮した情報推薦方式の研究に取り組んでいる[3][4]．例として，普段は大阪で生活をしている利用者を考える．ある日，この利用者が観光で三重県松阪市に来ていたとする．そして，食事をとるために携帯電話からローカルサーチサービスを利用して，松阪駅の近くにある飲食店と，その店が扱っている料理を検索した．すると，松阪駅から近くにあるさまざまな飲食店が検索されたが，中には，「マクドナルド」の「てりやきバーガー」や「ガスト」の「ステーキハンバーグ」など，大阪を含め，どこでも利用できるようなものも含まれていた．しかし，この利用者にとってはせっかくの観光であるため，これら全国チェーン店のように，どこでも利用できるような店には魅力を感じない．この利用者の地元である大阪では利用できず，現地である松阪でしか利用できないような料理（たとえば，「松阪牛」）を提供している店が検索された方が，この利用者にとっては魅力的であるといえる．

また，「松阪牛」のように「松阪 ⇒ 松阪牛」というように，その土地の名物を容易に連想できるものであれば，キーワード検索により検索可能であるが，あまりメジャーでない土地に観光に行ったときには，何が名物であるかも分からない．したがって，現地でしか利用できないような料理などに関するキーワードを自動的に抽出し，それを提示することで，利用者に気付きを与える必要がある

我々の先行研究[3][4]では，利用者にとって地元では利用できないが，旅行先や出張先など現地では利用で

きるようなスポット（飲食店や娯楽施設，観光施設など）を，地域限定性が高いスポットとし，この地域限定性を考慮した情報推薦方式を提案した．またその有用性を検証するため，奈良および松阪（三重），近江八幡（滋賀）を現地とした定性分析を行った．結果，例えば松阪では，例に挙げたように「松阪牛」に関連する語句が多く抽出されたり，奈良では「春日大社」や「東大寺」など奈良の地域性を反映させるような語句が抽出された．一方で，結果の一部には，奈良を現地としたときでさえ,「ダイニングバーオープン」や「フレンチテラス」などといった一般的な語句や単なる店舗名や駅名などを表す語句もみられた．これらの語句は，地域限定性を目指した推薦においては利用者にとってノイズとなるため，そのノイズとなるような語句を排除する方法について検討する必要がある．

そこで，本研究では，我々の先行研究[3][4]の結果を踏まえ，地域限定性を目指した推薦において，どのような語句が利用者にとって有用であるか，またノイズとなるかについて分析し，抽出語句からノイズ語句を排除する方法について検討する．

本稿の構成は以下のとおりである．第 2 章では，関連研究および関連事項を取り上げ，これらと本研究との違いについて述べる．第 3 章では，我々の先行研究において提案した地域限定性を考慮した情報推薦方式の概要を説明する．第 4 章では，先行研究の実験結果を引用し，その分析を踏まえ，ノイズ語句を排除する方法について説明する.第 5 章では,本稿をまとめる．

2. 関連研究

手塚ら[5]は，ウェブページやオブジェクト（「紅葉」や「うどん」など）が持つ「地域性」を推定する手法を提案している．たとえば,「紅葉」が有名な場所を調べたいとき,「紅葉」というオブジェクト名を入力することで, 最も「紅葉」と関連の深い地域を表示させる．つまり，オブジェクト名から関連する地域を取得するというものである．これに対し，本研究では，逆に地域からその地域に限定的な語句を抽出することを目指している．

Tarumi ら[6][7]は，時空間限定情報を扱うシステム SpaceTag を提案している．SpaceTag は，時空間限定でアクセス可能な仮想オブジェクトであり,特定の場所, 特定の時間でのみアクセスできるテキスト，画像，音声，プログラムなどの任意のオブジェクトである．SpaceTag は，企業や公的機関，一般ユーザによって作成される．したがって，作成者の主観や意図が大きく含まれる．これに対し，本研究では，実空間上の位置に関連付けられた膨大なスポット情報の中から自動的に地域に限定的な語句を抽出するものである．

また, Web 上には多くのご当地グルメサイトが存在する．このサイトを利用することで確実にご当地グルメの情報を得られるが，やはりサイト作成者の主観が入る．また，有名な観光地などにおいては，情報も充実しているが，あまり知られていないような土地においては，このようなサイトが提供されていないこともある．これに対し，本研究での提案方式は，実空間上の位置に関連付けられたスポット情報さえあれば，あまり知られていない土地でも対応可能である．また，人手によらず，自動的に地域限定的な語句を抽出するため，サイト作成者が気付かなかったような語句も発見できる可能性がある

3. 地域限定性を考慮した情報推薦

本章では, 我々の先行研究[3][4]における提案方式である地域限定性を考慮した情報推薦方式の概要を説明する．

提案方式は，スポットの地域限定性に着目し，地域限定性の高いスポットを推薦するものである．本研究では，実空間に存在する飲食店や娯楽施設，観光施設などをスポットとよび，利用者にとって地元では利用できないが，現地では利用できるようなスポットを，地域限定性が高いスポットとしている．スポットに関する情報は，ぐるなび[8]などの情報サイトから提供されており，住所などの位置情報をはじめ，スポットに関するさまざまな情報が取得できる．提案方式では，スポットに関する情報として，

・スポット名
・位置情報（緯度・経度もしくは住所）
・テキスト情報（PR 文など）

を利用する．

地域限定性を考慮した情報推薦を行うために，提案方式では，以下の手順により地域限定性の高い語句抽出を行う．

i 地元スポットおよび現地スポットの取得
ii スポットのテキスト情報からの語句抽出
iii 抽出語句の地域限定性スコアの算出

以下，各手順について簡単に述べる．

3.1 地元スポットおよび現地スポットの取得

ここでは，地元スポットおよび現地スポットについて述べる．利用者の地元に存在するスポットを地元スポット h_i とよび，地元スポット集合を，

$$H = \{h_1, h_2, \ldots, h_n\} \quad (1)$$

と表す．ここで，n は地元スポットの総数である．たとえば，利用者の地元が「大阪府」である場合，地元スポット集合 H には「大阪府」に存在する全スポットが含まれ，そのスポット数が n となる．

また，旅行先や出張先などの現地に存在するスポットを現地スポット l_j とよぶ．現地において基点となる位置を中心に半径 r の範囲内に存在する現地スポット集合を，

$$L = \{l_1, l_2, \ldots, l_m\} \quad (2)$$

と表す．ここで，m は範囲内に存在する全スポットの数である．たとえば，旅行先が「松阪」で，基点を「松阪駅」とした場合，現地スポット集合 L には「松阪駅」から半径 r の範囲内に存在する全スポットが含まれ，そのスポット数が m となる．

3.2 スポットのテキスト情報からの語句抽出

現地スポット集合 L 内の各スポット l_j のテキスト情報に含まれる語句を抽出する．語句抽出には，形態素解析器である茶筌[11]を利用する．ここで，語句抽出対象は名詞のみとした．また，語句の特徴が失われることを防ぐため，たとえば「松阪」＋「牛」を「松阪牛」とするように，連続する名詞は連結させる．さらに，「松阪市の老舗料亭」のように，「助詞-連体化」により連結されうる名詞句を一つの語句として扱う．

スポット l_j に対して，抽出された語句の集合を，

$$W_j = \{w_{j1}, w_{j2}, \ldots\} \quad (3)$$

とする．

3.3 抽出語句の地域限定性スコアの算出

抽出された各語句 w_{jk} が，どの程度その地域に限定的なものであるかを調べる．これを調べるため，文書検索によく用いられる IDF（文書頻度の逆数）[9]を適用する．

この考えに基づき，現地スポットのテキスト情報に出現する各語句 w_{jk} が，地元スポット集合に対し，どの程度限定的なものかを表す尺度として，限定性 ν_{jk} を定義する．式（4）において，文書をスポットのテキスト情報と置き換えると，この限定性 ν_{jk} は次式のように求められる．

$$\nu_{jk} = \log \frac{n+1}{n_{jk}+1} \quad (4)$$

ここで，n は地元スポット集合 H に含まれるスポット数（3.1 節参照）であり，n_{jk} は，地元スポット集合 H のうち，語句 ν_{jk} をテキスト情報に含むスポットの数である．この限定性を用いることで，利用者にとって地元のスポットには現れないが，現地のスポットにしか現れないような限定的な語句を抽出することができる．

また，現地の地域性を反映させるために，限定性 ν_{jk} に対し，地域関連重み γ_{jk} を付加する．地域関連重み γ_{jk} は，語句 w_{jk} が，どの程度，現地の地域に関連しているかを表す重みである．すなわち，「松阪」における「松阪牛」のように現地との関連が強い語句には高い重みを与え，「クーポンのサラダバイキング付」のように現地との関連が弱い語句には低い重みを与える．

この地域関連重みを考慮するために，Web 上での単語の共起頻度に基づいた単語類似度を表す指標である WebPMI[10]を用いる．単語 p および q の WebPMI は次式で表される．

$$\mathrm{WebPMI}(p,q) = \begin{cases} 0 \\ \log \dfrac{H(p \cap q)/N}{H(p)/N \;\; H(q)/N} \end{cases} \quad (5)$$

ここで，$H(p)$, $H(q)$, $H(p \cap q)$ は，それぞれ，「p」，「q」，「p + q」をクエリとして検索を行ったときのヒット件数である．N は，検索エンジンが持つ全文書数である．また，c は低頻度語によるノイズを避けるために用いられる閾値である．

語句 w_{jk} と現地の地域名 local との WebPMI($w_{jk,}$ local) を求めることによって，語句 w_{jk} の地域関連重み γ_{jk} を算出する．つまり，γ_{jk} は次式のように定義される．

$$\gamma_{jk} = \mathrm{WebPMI}(w_{jk,}\ \mathrm{local}) \quad (6)$$

ここで，現地の地域名の抽出には逆ジオコーディング 11) を用いる．基点位置の緯度・経度から逆ジオコーディングにより取得された市区町村名（「松阪市」など）を local とする．

以上の限定性 ν_{jk} および地域関連重み γ_{jk} の二つの尺度を考慮して，語句 w_{jk} の地域限定性スコア s_{jk} を次式から求める．

$$s_{jk} = \nu^*_{jk} \times \gamma^*_{jk} \quad (7)$$

$$\nu^*_{jk} = \frac{\nu_{jk} - \min_k \nu_{jk}}{\max_k \nu_{jk} - \min_k \nu_{jk}} \quad (8)$$

$$\gamma^*_{jk} = \frac{\gamma_{jk} - \min_k \gamma_{jk}}{\max_k \gamma_{jk} - \min_k \gamma_{jk}} \quad (9)$$

ただし，ν^*_{jk} および γ^*_{jk} は，それぞれスポット l_j において値を[0; 1] に正規化したものである．

4. 抽出語句の傾向分析とノイズ語句の排除

3 章で述べた提案手法の有用性評価に関し，先行研究では，地元を「大阪府」とし，現地を「奈良公園」として基礎実験を行った．提案手法に関して定性的に有用性を確認した一方で，抽出語句に関して，いくつかのノイズとなる語句が含まれていた．本論文では，より

多くの地域を実験対象にし，提案手法による抽出語句の分類を行ったうえで，ノイズとなるような語句の傾向を分析する．その分析結果を踏まえ，ノイズ語句の排除方針について検討する．

4.1 データセット

提案方式は，飲食店や娯楽施設，観光施設などに対して適用し得る手法であるが，本実験では，スポットのジャンルとして飲食店を対象とした．飲食店データは，一般向けに公開されているグルメ情報サイトである，ぐるなび[8]から取得した．本実験では，ぐるなびが提供している API[13]により，各飲食店の

・飲食店名
・位置情報（緯度・経度）
・テキスト情報（PR 文（短）および PR 文（長））

を取得した．

本実験においては，地元スポットを「大阪府」とした．すなわち，ぐるなび API により取得可能な大阪府内の全飲食店データの集合を地元スポット集合とした．

また，現地として，名古屋駅（緯度：35.170694，経度：136.881637）と京都府庁（緯度：35.021247，経度：135.755597）を選び，それぞれを基点とした．各基点から半径 3,000m の範囲内に存在する全飲食店データの集合を，各基点における現地スポット集合とした．

4.2 先行研究の実験結果における抽出語句の傾向

まず，先行研究の実験結果を表 1 に示し，抽出語句の傾向をみる．先行研究の実験では，地元を「大阪府」とし，現地を「奈良公園」とした．3.2 節で述べた方法にしたがい，現地スポット集合内の各飲食店データのテキスト情報から語句抽出を行った．抽出された語句は 483 個であった（ただし重複語句は排除した）．

抽出された各語句に対し，地域限定性スコアを算出し，このスコアにしたがって，語句のランキングを行った．ただし，式 (6) において $c = 5$ とした．ここで，$H(p \cap q) \leqq c$ となった語句は 47 個であった．これらはランク外とし，残りの 436 個についてランキングを行った．その結果のうち上位 20 件を，**表 1** に示す．

奈良県などの観光名所が多い地域は比較的にスポットデータが多いことから，上位ランクにその地域の有名なスポットを表す語句が多く含まれるという傾向がみられた．しかし，「ひがしむき商店街内」（16 位），「近鉄奈良駅」（17 位）などは，単にそのスポットの所在地を示している語句である．また，「フレンチテラス」（11 位）や「テラススタイル」（23 位）などは現地である奈良特有のものであるとはいえない．したがって，地域限定的なスポットを推薦する提案手法の目的からすると，これらの語句が提示されることが利用者にとって有用であるとはいえない．

表 1 地域限定性スコアに基づくランキング結果

ランク	語句	スコア
1	春日山の大自然ただ中	1.0000
2	名立たる社寺の中心	0.9800
3	秋吉博国	0.9756
4	古都の宿	0.9687
5	底の鰹	0.9655
6	春日大社	0.9158
7	奈良公園	0.9151
8	特製の鰹	0.9094
9	大和肉鶏	0.9032
10	ダイニングバーオープン	0.8980
11	フレンチテラス	0.8970
12	木質調	0.8910
13	吉野本葛	0.8852
14	東大寺	0.8781
15	奈良の町	0.8723
16	ひがしむき商店街内	0.8592
17	近鉄奈良駅	0.8468
18	こだわりのかつお	0.8440
19	百楽	0.8196
20	春日奥山原始林	0.8118

本節では，名古屋駅と京都府庁を現地の基点としたときの抽出語句の傾向をみる．**表 2 (a)** および **(b)** に，それぞれを現地としたときの抽出語句のランキング結果を示す．

表 2 (a) より，名古屋駅を基点とした場合，地域限

表 2(a) 基点を名古屋駅にした時のランキング結果

ランク	語句	スコア
1	久屋大通公園	0.98499382
2	だい	0.91763798
3	丸の内	0.89763191
4	貸切パーティー会場	0.89519274
5	伏見	0.88605918
6	名古屋駅	0.88414238
7	栄	0.88414238
8	晩	0.87767497
9	拠点	0.87459685
10	丸の内徒歩	0.87392893
11	南風	0.87208703
12	居酒屋のメニュー	0.87205682
13	社長気分	0.87149744
14	山ちゃん	0.87089055
15	個室ば	0.87027783
16	名古屋の中心街	0.86967093
17	桜通	0.86967093
18	ご人数分	0.86788467
19	コースの始まり	0.86786174
20	組合せ	0.86610319

表 2(b) 基点を京都府庁にした時のランキング結果

ランク	語句	スコア
1	二つのテーブル席	1
2	いごこち	0.49308203
3	界わい	0.46613067
4	オトク	0.4590478
5	抜群の四条寺町	0.43110107
6	レディースセット	0.42983085
7	京都最大級	0.42758263
8	お気にいり	0.42582759
9	京都徒歩	0.41813048
10	鳥取県	0.41694367
11	スペシャルディナーの催し	0.41530734
12	京都の中心街	0.41478051
13	ツアープロ	0.4099812
14	抜群の四条烏丸	0.40575819
15	方々の憩いのひととき	0.39891349
16	関西初	0.3947359
17	おためし	0.3904437
18	鱧なべ	0.38950326
19	会社の仲間	0.38510192
20	だい	0.38309741

定的な語句として，「久屋大通公園」（1 位）が抽出できた．しかし，例えば，店の名前である「山ちゃん」（14 位）や「丸の内」（3 位）などは，地域限定的なスポットを発見するという本研究の目的では，利用者にとっては提示されてもあまり有用ではないといえる．また「栄」（7 位）などの都市名や「名古屋駅」（6 位）などは利用者にとっては自明であるため，これらも提示されて有用であるとはいえない．

表 2 (b) より，京都府庁を基点とした場合，地域限定性の特徴を表す語句として，「鱧なべ」（18 位）の語句は抽出できた．しかし，上記と同様に，「二つのテーブル席」（1 位）などは，一般的な語句であるため提示されても有用でないといえる．

以上のように，提示されても有用でないような語句を本研究ではノイズ語句として扱う．抽出されたノイズ語句の内容に着目すると，ノイズ語句は大きく以下のように分類される．

a) 「連体助詞」（「の」）により連結された語句
b) 店名や都市名，駅名などを表した語句
c) 一般的な文書に頻出する語句

以下，各分類について考察する．

4.2.1「連体助詞」により連結された語句に対する考察

分類 a) に属する語句として，表 3 (a) および (b) に示す語句が抽出された．

表 3 (a) にみられるように，「奥三河鶏の香草焼」（31 位）に関しては，名古屋特有の料理であることから有用であるといえる．また，表 6 における「えびのてんぷら」（43 位）に関しては，全国で利用できる料理かもしれないが，名古屋では特別な料理である可能性もある．しかしながら，一方では，「居酒屋のメニュー」（**表 3 (a)** 12 位）や「名古屋の中心街」（**表 3 (a)** 16 位），「二つのテーブル席」（**表 3 (b)** 1 位），「抜群の四条寺町」（**表 3 (b)** 2 位）などの語句は，利用者にとって有用な語句であるとはいえない．

3 章で述べた方法では，語句の特徴を失わせないために「連体助詞」により連結される語句を一つの語句として抽出を行った．しかし，全体的にみると「連体

表 3 (a) 分類 a) に属する語句（基点：名古屋駅）

語句	ランク
居酒屋のメニュー	12
名古屋の中心街	16
コースの始まり	19
沖縄の風	22
お勧めのコース	24
地下鉄名城線の久屋大通り	29
奥三河鶏の香草焼	31
名古屋の丸の内	32
当店のHP	33
会社の帰り	36
雰囲気の東区	41
当店の特徴	42
近隣のお客様	44
こだわりの宴会	45
ソファーの席	49
大人のイタリアン	58
世界の山ちゃん	59
バーの落ち着き	72
オーナーの自信作	75
名古屋の老舗	81

表 3 (b) 分類 a) に属する語句（基点：京都府庁）

語句	ランク
二つのテーブル席	1
抜群の四条寺町	5
スペシャルディナーの催し	11
京都の中心街	12
抜群の四条烏丸	14
方々の憩いのひととき	15
会社の仲間	19
魅力の京酒場	22
老舗の本格派	25
茶懐石の流れ	28
純和風の個室	32
えびの天ぷら	43
個室季節の流れ	59
コースの中身	64
タイプのお部屋	82
しゃぶしゃぶの店	161
ゲストのニーズ	181
宴会の個室	195
最新のゴルフクラブ	197
人気のゲームメニュー	201

助詞」により連結される語句には一般的なものも多く，本研究の目的からはあまり有用な語句であるとはいえないことが分かる．これは，一般語句同士が連結されることによって，地元スポットのテキスト情報集合に含まれない語句となり，限定性の値が高くなってしまったことが原因である．

そこで，本研究での対策としては，「連体助詞」による語句の連結を行わずに，それぞれを独立した語句として扱う．

4.2.2 店名や都市名，駅名を表した語句に対する考察

分類 b) に属する語句として，**表 4 (a)** および **(b)** に示す語句が抽出された．

表 4 (a) 分類 b) に属する語句（基点：名古屋駅）

語句	ランク
丸の内	3
伏見	5
名古屋駅	6
栄	7
南風	11
山ちゃん	14
桜通	17
沖縄の風	22
地下鉄名城線の久屋大通り	29
久屋大通駅	34
久屋大通り駅	39
丸の内駅	40
東桜	46
世界の山ちゃん	59
月あかり	116

表 4 (b) 分類 b) に属する語句（基点：京都府庁）

語句	ランク
鳥取県	10
東山	39
新橋	60
吉田	70
明石	74
大宮店	80
江戸川	86
足立	93
築地	96
坂本	110
滋賀県	118
富山県	136
大宮駅	137
松原	142
京都駅	153

表 4 (a) にみられるように，「山ちゃん」(14 位)や「沖縄の風」（22 位）は，飲食店名である．そもそも飲食店名は固有のものが多く，限定的な語句として抽出されやすい．しかし，飲食店名がその店の特徴を表しているとは限らない為，飲食店名を提示することは，利用者にとって有用でないと考えられる．

また，「名古屋駅」（6 位）や「久哉大通駅」（34 位），「伏見」（5 位），「栄」（7 位）などの駅名や都市名などはその地域と関連が深い為，地域関連重みの値が高くなり，上位にランキングされた．表 4 (b) でも同様に，多くの都市名が抽出された．さらに，県名である「鳥取県」（10 位）が上位にランキングされた．しかし，提案の推薦方式では，利用者が現地を指定して，現地周辺の地図上にスポットとその地域限定的な特徴語句を提示することを想定しているため，スポットの位置に関しては利用者にとっては自明である．そのため，スポットの位置を説明するためだけに用いられているランドマーク名などを提示することは，有用であるとはいえない．

よって，飲食店名と駅名，都市名に関してはノイズ語句として扱う．

4.2.3 一般的な文書に頻出する語句に対する考察

一般語を判断する基準として，3.3 節で説明した IDF を適用する．Web ページ全体を対象とした IDF 値に基づいて，地域 A と地域 B それぞれの基礎実験の結果から上位２００位までの抽出語句をランキングした結果を**表 5(a)，(b)**に示す．また，地域 B においては，テキスト情報が多い為，上位 30 件まで判断を行い，下位 20 件の判断を行った結果を**表 5(a)，(b)**に示す．ここでは，特徴的な語句を 1 とし，一般的な語句を 0 として分類した．また，有用的な語句の判断をする為に，判断を行った．

表 5(a)，(b)，表 6(a)，(b)にみられるように，IDF 値が 11.1.を境目に特徴的な語句と一般な語句に分かれていることがわかる．また，「浜」，「からだ」など一般的に使われる語句が下位にきていることが確認できた．ここでいう分類 c)は，Web ページ全体を対象とした I

表 5(a) 上位 20 件の語句の判断（基点：名古屋駅）

上位20件			
語句	IDF値	分類	判断
奥三河鶏の香草焼	16.82542604	1	有用
アメニティ泉パーキング	14.25589477	1	無用
地階の隠れ家	14.19493407	1	無用
四国愛媛の名物今治焼き	14.11241305	1	有用
栄の北寄り	12.63058351	1	無用
小ジョッキ通常	12.61495819	1	無用
コースのお客様ダーツ無料特典	12.49312284	1	無用
大ジョッキ通常	11.9363121	1	無用
夏の旬魚コース等	11.82479083	1	無用
柳橋市場	11.7381983	1	無用
栄寄り	11.68001634	1	無用
中ジョッキ通常	11.63745673	1	無用
地下鉄名城線の久屋大通り	11.55357524	1	無用
丸の内の名店	11.47618858	1	無用
フラリ	11.36870267	0	無用
鉄	9.852665928	0	無用
ワゴン	9.852665928	0	無用
本気	9.845697258	0	無用
伏見	9.838776816	0	無用
ワインバー	9.838776816	0	無用

表 5(a) 下位 20 件の語句の判断（基点：名古屋駅）

下位20件			
語句	IDF値	分類	判断
店内スペース	9.740336743	0	無用
一軒	9.740336743	0	無用
食材の味	9.740336743	0	無用
居酒屋のメニュー	9.734106193	0	無用
ご人数分	9.734106193	0	無用
前餅	9.734106193	0	無用
中華ダイニング	9.734106193	0	無用
デザート付	9.734106193	0	無用
レッドオレンジ	9.734106193	0	無用
町通	9.734106193	0	無用
お勧めのコース	9.727914223	0	無用
お酒達	9.727914223	0	無用
食空間	9.727914223	0	無用
中京テレビ	9.727914223	0	無用
組合せ	9.721760357	0	無用
ご家族様	9.709565084	0	無用
空間美	9.662218963	0	無用
会社員	9.386164929	0	無用
浜	8.306741421	0	無用
晩	8.116010218	0	無用

表 6(a) 上位 20 件の語句の判断（基点：京都府庁）

上位30件			
語句	IDF値	分類	判断
ゆばんさい各種	15.30327091	1	無用
抜群の四条寺町	14.20574499	1	無用
スペシャルディナーの催し	14.08314266	1	無用
お忍び仕様	13.69058096	1	無用
山科なす	13.41432758	1	有用
界わい	13.19414407	1	無用
抜群の四条烏丸	12.76652822	1	無用
方々の憩いのひととき	12.68252825	1	無用
魅力の京酒場	12.61828598	1	無用
せみしぐれ	12.28978191	1	無用
いごこち	12.13448903	1	無用
水たき	11.97643209	1	無用
存知	11.65426385	1	無用
純和風の個室	11.50496754	1	無用
茶懐石の流れ	11.46210384	1	無用
賀茂なす	11.3220379	1	有用
レディースセット	11.31296068	1	無用
加茂なす	11.24035987	1	有用
ソロバン	11.23199162	1	無用
個室季節の流れ	11.13415937	1	無用
木藤	11.12413422	0	無用
ツアープロ	11.10928247	0	無用
お気にいり	11.08022473	0	無用
宴会スペシャルプラン	11.0636563	0	無用
牛ホホ肉	10.99560284	0	無用
伊万里焼	10.96111667	1	有用
全室サイバー	10.8508951	0	無用
焼肉盛合せ券	10.79912754	0	無用
ロフト席	10.69810072	0	無用
老舗の本格派	10.59306599	0	無用

DF 値が 11.1 以下であり，分類 a)，b)を除いた語句を分類 c)とし，一般語句と定義した．

一般語句と定義されたものにも限定性を含まれる語句が多く存在しているが，地域関連重み γ_{jk} を加えることによって地域限定性の高い語句のみ上位に抽出できる．また，IDF 値と地域関連重み γ_{jk} の低いスコアや IDF 値が高いが，地域関連重み γ_{jk} の低いスコアはノイズ語句となる．

表 6(b) 下位 20 件の語句の判断（基点：京都府庁）

下位20件			
語句	IDF値	分類	判断
フレンチランチ	9.752915525	0	無用
京都タワー	9.752915525	0	無用
京都の中心街	9.746606356	0	無用
ゴールデンウイーク	9.746606356	0	無用
各種メニュー	9.746606356	0	無用
関西初	9.740336743	0	無用
二つのテーブル席	9.727914223	0	無用
いのしし肉	9.727914223	0	無用
鳥取県	9.721760357	0	無用
会社の仲間	9.721760357	0	無用
京都初	9.721760357	0	無用
コースの中身	9.721760357	0	無用
タイプのお部屋	9.721760357	0	無用
バイキング付	9.633727008	0	無用
フラワー	9.579068595	0	無用
かい	9.532306829	0	無用
富山県	8.891254761	0	無用
滋賀県	6.428426052	0	無用
チョコレート	5.71164243	0	無用
からだ	5.366153286	0	無用

4.3 ノイズ語句の除去

4.2 節の分析結果より，以前の地域限定性を考慮した情報推薦方式に以下のように追加・変更し，ノイズの除去を行う．

- Web ページ全体を対象とした IDF 値を限定性スコアに追加する．
- 飲食店名，駅名を上位スコアカラ取り除く為，地元スポットの集合に現地スポットの飲食店名，都市名，駅名の情報を追加する．
- 連結語句は「名詞」＋「名詞」のみに限定する．

ノイズの排除を行った結果の一部を地域 A と地域 B それぞれ**表 7(a)，(b)**に示す．

表 7(a) ノイズ排除を行った結果(基点：名古屋駅)

ランク	語句	地域関連重み γ_{jk}	限定性 s_{jk}	Webページ全体のIDF値	スコア
1	窓際プレミアムシート	1	1	0.644477025	0.644477025
2	名取観男	1	0.705779671	0.904275832	0.6382195
3	純系名古屋コーチンガラ白湯美人鍋	1	0.648894034	0.855661932	0.555233923
4	バイナ	1	0.919387994	0.601891285	0.553371621
5	ホームメイドチキン居酒屋さちこ名古屋伏見	1	0.693657435	0.753200453	0.522463094
6	センチュリー豊田ビル元祖手羽先	1	0.652257624	0.784464672	0.511673063
7	核豆笹葉包み寿しすずき黄味焼	1	0.546266204	0.85370276	0.466348966
8	各種名古屋名物	1	0.881924131	0.528127255	0.465768171
9	知多牛づくし	1	0.669632156	0.689007676	0.461381696
10	居酒屋グルメ卸関座	1	0.68151164	0.675577453	0.460413898
11	大須観音徒歩	1	0.823322148	0.54973069	0.452605452
12	おひろめ	1	0.793873714	0.564651074	0.448261645
13	寛福バリューブラン	1	0.592400611	0.742626775	0.439932555
14	地元名柄豚美濃	1	0.591412242	0.741775853	0.43869532
15	岡崎八丁味噌のど	1	0.598356128	0.700860718	0.419364306
16	飛騨牛ちりとり焼	1	0.595921861	0.703720867	0.419362648
17	名古屋名物極上串かつ	1	0.61037974	0.677823728	0.413607795
18	盛北ルモンスペシャル	1	0.744323409	0.553807121	0.412211604
19	各種薩美会席	1	0.809592289	0.507394752	0.410782879
20	炭焼割烹わらじ	1	0.576783432	0.702197157	0.405015686
21	夏ばて	1	0.792008677	0.506776842	0.401371656
22	地酒	0.917637979	0.875658939	0.497122653	0.399456892
23	込ディナーコース	1	0.783656866	0.504873115	0.395647283
24	おもち	1	0.873875154	0.449548177	0.392848982
25	般膳	1	0.782914163	0.500387596	0.391760536
26	大須酒場	1	0.649411994	0.585544987	0.380259944
27	三河鶏鍋コース	1	0.652829048	0.576710666	0.376493475
28	宮崎名物炭焼日向鶏	1	0.524072622	0.718260713	0.376420776
29	大好評甘口	1	0.67256474	0.558279918	0.375479388
30	全商品和牛	1	0.751866909	0.495895017	0.372847054
31	蔵元盛田	1	0.557058498	0.667519328	0.371847314
32	三河赤鶏	1	0.654669089	0.562542296	0.368279053
33	三河牛ロース	1	0.587629069	0.625541222	0.367586206
34	愛知県知事賞	1	0.705779671	0.520224932	0.367164181
35	豊浜漁港	1	0.594562494	0.614545857	0.365385917
36	知多産	1	0.665675157	0.53838475	0.355360548
37	めん三河地	1	0.624740881	0.566112581	0.353673672
38	なごや純米樽秋	1	0.527049704	0.668538564	0.352353052
39	飛騨牛鉄板焼ぎ	1	0.60294233	0.582319961	0.351105354
40	富山新湊漁港	1	0.563025643	0.622663186	0.35057534

表 7(a)，(b)をみると，「純系名古屋コーチンガラ白湯美人鍋」(**表 7(a)** 3 位)や「知多牛づくし」(**表 7(a)** 9 位)，

表 7(b) ノイズ除去を行った結果（基点：京都府庁）

ランク	語句	地域関連重み τ_{jk}	限定性 ν_{jk}	Webページ全体のIDF値	スコア
1	界わい	1	0.466130668	0.71752831 3	0.334461952
2	いごこち	1	0.493082034	0.658501468	0.324695243
3	ゆばさい各種	1	0.346150606	0.829933811	0.287282091
4	お忍び仕様	1	0.345383661	0.742769894	0.256540586
5	山科なす	1	0.352092951	0.727820879	0.256260601
6	蓮月尼	1	0.297291118	0.828807721	0.246397174
7	太秦石田様	1	0.294869649	0.77519671	0.228581982
8	水たき	1	0.347981568	0.648669421	0.225725002
9	茶懐石	1	0.362393573	0.622013188	0.225413581
10	ソロバン	1	0.36635621	0.609525704	0.223303527
11	全室サイバー	1	0.375968252	0.589779531	0.221738379
12	加茂なす	1	0.362300649	0.608332976	0.220398432
13	せみしぐれ	1	0.330005896	0.666928736	0.220090415
14	京都保津峡鵜	1	0.267134508	0.816237429	0.218045184
15	祇園乙部	1	0.259603507	0.838240775	0.217610245
16	鱧なべ	1	0.389503263	0.556017205	0.216570516
17	存知	1	0.341900802	0.632441122	0.216232127
18	新種伝統建物都	1	0.268874743	0.794098245	0.213512962
19	桃屋高瀬川船入	1	0.242664188	0.86143005	0.209038223
20	関西初	1	0.394735895	0.528918406	0.20878306
21	大丸櫃達歩	1	0.293456007	0.705684362	0.207067315
22	おためし	1	0.390443699	0.530301078	0.207052714
23	清水吉次郎邸	1	0.221208541	0.935883625	0.207025451
24	南禅寺	1	0.249137862	0.823736942	0.205224061
25	だい	0.855622686	0.447741067	0.533547244	0.204400569
26	焼肉盛合せ券	1	0.348757751	0.586035499	0.204384423
27	木藤	1	0.336576977	0.605182032	0.203693039
28	椀盛司	1	0.217033697	0.935883625	0.203118283
29	なか川	1	0.31275086	0.647747014	0.202583436
30	つけもの	1	0.35352693	0.571311094	0.201973857
31	博多風水たき	1	0.241967722	0.8324416 4	0.201424007
32	我流たくみ	1	0.294916095	0.681309893	0.200929253
33	賀茂なす	1	0.32524584	0.614412239	0.199835025
34	スッポン鍋	1	0.376321956	0.529260785	0.199172454
35	岩手十文字鶏	1	0.254778576	0.774878428	0.197422423
36	山陰竹野浜漁港	1	0.234017494	0.841174837	0.196849628
37	わんこそば	1	0.284991851	0.686720272	0.195709681
38	お米京北町	1	0.294841631	0.663562266	0.19564578
39	京都祇園像員	1	0.217451122	0.894689319	0.194551196
40	坦種	1	0.310572535	0.625743668	0.194338787

「山科なす」**(表 7(b)** 5 位)，「加茂なす」**(表 7(b)** 12 位)などかなりの限定性語句が上位ランクに存在し，一般語句を下位ランクに下げることができた.「名取龍男」**(表 7(a)** 2 位)は，名古屋で有名な料理人であるが，料理人の名前を推薦しても有用的ではない．こういった限定性のある語句に対しても考えていく必要がある．結果として，全体的には大量のノイズ語句を排除することができたが，多くのノイズ語句はかなり含まれる．これは「窓際プレミアムシート」**(表 7(a)** 1 位)や「ソロバン」**(表 7(b)** 10 位)など地元のテキスト情報のデータが不足したことで，ノイズ語句に対する限定性 ν_{jk} の値が高くなってしまったことや，形態素解析する際に，全角スペース部分が排除されてしまい,「ホームメイドチキン居酒屋さちこ名古屋伏見」**(表 7(a)** 5 位)など意味のない名詞同士の連結がされてしまい，必然的に IDF 値が高くなってしまったことが原因である．名詞同士の連結に関しては，今後検討が必要になってくる．

また，特殊なノイズ語句として，形態素解析でうまく分割できなかった語句の IDF 値が高くなってしまうというケースも見られた．例えば，地域 A で実験を行った際に，56 位に「国産うなぎひつ」などが抽出されたが，これは「国産うなぎのひつまぶし」がうまく分割できなかったことが原因である．特殊なノイズ語句に対しての検討もしていく必要がある．

5. まとめ

先行研究では利用者にとって地元では利用できないが，旅行先や出張先など現地でしか利用できないようなスポットを地域限定性の高いスポットとし，この地域限定性を考慮した情報推薦方式を提案した．先行研究の課題であった利用者にとって有用的な語句ではないノイズ語句を排除することが本研究の目的である．ノイズ語句には様々な種別に分類されることを確認した．そして，分類した各項目において，ノイズ語句の分析と傾向を行いノイズ語句の排除を行った．今後の展望として，情報推薦システムの構築を目指し，以下の点について検討していく．

- 再現率。適合率などの評価尺度に基づいた定量的分析を行うとともに，従来の語句抽出との比較による有効性の評価を行う．
- 利用者満足度の視点から提案方式の有用性を検証する．
- ノイズ語句として抽出された「連体助詞」（「の」）により連結された語句に対し，有用的である語句を抽出への対応について検討する．
- 限定性(記号)の精度をより上げるために，テキストの追加・選定を行う．
- 今回は，現地スポットを変更せずに，ノイズ語句の分析を行ったが，現地スポットを変えた時のノイズ語句の傾向の分析を行う．

参 考 文 献

[1] Google マップ. http://maps.google.co.jp/.

[2] Yahoo!地図. http://map.yahoo.co.jp/.

[3] 奥 健太，服部 文夫: 地域限定性を考慮した情報推薦方式に関する基礎検討, Web とデータベースに関するフォーラム，情報処理学会シンポジウムシリーズ, Vol.2009, No.3, pp.1A-1, 2009.

[4] Kenta Oku and Fumio Hattori: Basic Study on a Recommendation Method Considering Region -restrictedness of Spots, DASFAA 2010, International Workshops: SNSMW, LNCS, Vol.6193, pp.353-364, 2010.

[5] 手塚 太郎，近藤 浩之，田中 克己．混合ガウス分布を用いたウェブコンテンツの地域性推定とオブジェクトレベルローカルサーチ．情報処理学会論文誌：データベース, Vol.1, No.1, pp. 13–25, 2008.

[6] H. Tarumi, K. Morishita, and Y. Kambayashi. Public applications of spacetag and their impacts, digital cities: Technologies, experiences and future perspectives. LNCS, Vol.1765, pp.350–363, 2000.

[7] 森下 健，中尾 恵，垂水 浩幸，上林 弥彦．時空間限定オブジェクトシステム: Spacetag プロトタイプシステムの設計と実装．情報処理学会論文誌, Vol.41, No.10, pp. 2689–2697, 2000.

[8] ぐるなび. http://www.gnavi.co.jp/.

[9] 形態素解析システム茶筌. http://chasen.naist.jp/hiki/ChaSen/.

[10] 北研二，津田和彦，獅々堀正幹．情報検索アルゴリズム．共立出版, 2002.

[11] D.Bollegala, Y.Matsuo, and M.Ishizuka. Measuring semantic similarity between words using web search engines. In WWW2007, 2007.

[12] Google maps api. http://code.google.com/intl/ja/apis/maps/

[13] ぐるなび web サービス. http://api.gnavi.co.jp/api/service.htm.