集合知に基づく観光地推薦システムの構築
Sightseeing Location Recommendation System Based on Collective Intelligence

嶋田 和孝，上原 尚，遠藤 勉
九州工業大学 情報工学部 知能情報工学科
福岡県飯塚市川津 680-4
E-mail: shimada@pluto.ai.kyutech.ac.jp

Kazutaka Shimada, Hisashi Uehara, Tsutomu Endo
Department of Artificial Intelligence
Kyushu Institute of Technology
680-4 Kawazu Iizuka Fukuoka 820-8502 Japan

概要

本論文では，Web から獲得される様々な情報に基づき特徴化された観光地情報を利用し，その類似性を評価することで，観光地を推薦するシステムを提案する．本システムでは，入力をユーザのお気に入りの観光地とし，それをユーザの嗜好情報であると考える．我々は（1）Yahoo 知恵袋・ブログ上での共起キーワード，（2）時系列分布，（3）Yahoo 知恵袋上でのカテゴリ構造，（4）観光地周辺施設，（5）地図画像の 5 つの情報に着目し，各観光地に対する特徴ベクトルを生成する．これらの特徴ベクトルに対して，コサイン尺度に基づき類似度を算出し，類似性の高い観光地を推薦する．さらに観光地間の類似箇所や差異を明確に表現するため，情報を可視化する．実装した観光地推薦システムの動作例を検証し，その有効性を確認する．
Keywords: 集合知, Web，観光地推薦，類似度

Abstract

This paper describes a prototype system of sightseeing location recommendation. An input for the prototype system is a user's favorite location or facility. On the basis of the input, the system computes some similarity measures from sightseeing locations in our database. In the sysyem, we focus on five similarity measures; (1) co-occurrence keywords, (2) time sequence, (3) category information on Yahoo Chiebukuro, (4) surrounding area information and (5) map images. Our method visualizes the output of the recommendation for the comparison between the user's input and the system's output. We implement the prototype system as a Web application, and evaluate the effectiveness of it.
Keywords: Collective Intelligence, Web, Sightseeing recommendation, Similarity

1 はじめに

近年，スマートフォンやノートパソコンの普及により，容易かつ迅速に大量の情報を得ることができ，なおかつ情報発信も行えるようになってきている．特に，観光情報ではそれが著しく，各地の観光施設や公共団体から公開されているものもあれば，旅行経験者によるブログの投稿，Q&A サイトでのお薦め観光地の質問・回答など，多種多様な観光情報が存在する．しかし，そこから得られる情報量が膨大であり，どの観光地が自分の嗜好にあっているかなど，ユーザにとって正確な情報を素早く見つけ出すことは困難である．そのような背景から，観光に関する情報分析システムの開発が盛んに行われている [1, 2]．

本論文では，ユーザの嗜好を考慮した観光地推薦システムの構築を目的とする．観光地を推薦するシステムを考える場合，どのような環境を想定するかで戦略が異なる．たとえば，観光中の人に対して，次にどこに行くべきかを推薦する場合，移動可能な範囲でユーザの嗜好に合った場所を提示するべきであろう．同様に，これから

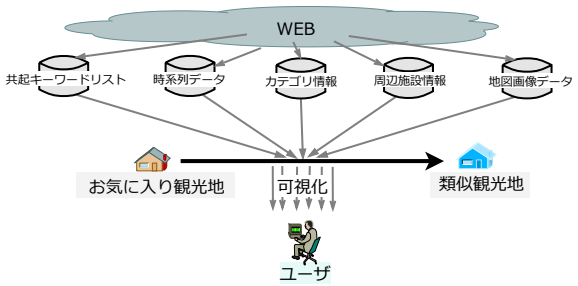


図 1　観光地推薦システムの概要図

行く観光のプランを考える場合にも，ユーザの嗜好を考慮して，行き先から候補を選ぶべきである．どちらの場合も，どのようにしてユーザの嗜好を捉えるかがシステムの入力として重要である．Kanazawa ら [3] は印象語などの分析に基づき，行きたい場所のイメージから連想される行き先推薦システムを提案している．本論文で提案するシステムでは，入力として具体的な 1 つの観光地を与える．これは，観光中の次の行き先を求めている場合では，今いる観光地にあたり，観光プランを考えている場合では，自分の知っている好きな観光地にあたる．この情報を与えることで，対象となる範囲の観光地との類似性を判断し，最適な観光地を推薦する枠組みの実現を目指す．

類似性を判断する場合に，何を基準に観光地間の類似度を測定するかが重要である．観光地に関する記事などのテキストデータを基に類似性を測ることが一般的であるが，対象とするすべての観光地に十分なテキストデータが存在するとは限らない．このような場合，有名な場所しか推薦されず，テキスト量の少ない知名度の低い観光地であればあるほど，情報不足により推薦されにくいという現象に陥る．この問題を解決するために，本論文では，集合知（Collective Intelligence）としての Web に着目する．ここで，集合知とは，多くの人による大量の情報を寄せ集めて，加工・集計した知的情報システムであると考える．複数の異なる情報源，異なる種類の情報を複合的に解釈し，類似度を算出することで，1 つの情報源のみに依存しない推薦システムの構築を目指す．

本システムの概要図を図 1 示す．Web からの情報源として，日本版 Wikipedia[*1]，Yahoo 知恵袋[*2]，ブログ，地図画像を用いる．観光地の特徴ベクトルは，(1) Yahoo 知恵袋・ブログ上での共起キーワード，(2) 時系列分布，(3) Yahoo 知恵袋上でのカテゴリ構造，(4) 観光地周辺施設，(5) 地図画像から生成する．本研究では，これらの特徴ベクトルから観光地間の類似度の測定を行い，類似度の高い観光地を推薦し，さらにどこが特徴的なのかを明確に表現するため，情報を可視化する．

本論文では，まず 2 節でシステムが利用する 5 つの観光情報について述べる．続いて，3 節で提案するシステムで利用するデータベースと類似度計算法および推薦のための戦略について述べる．4 節で作成した推薦システムの動作例を基にその有効性を検証し，システムへの改良のための議論をし，最後に 5 節でまとめる．

2　観光情報

本節では，構築する観光地推薦システムで利用する 5 つの観光情報について説明する．

2.1　共起キーワード

ユーザの入力したキーワードと各観光地に関連する単語と関連性は，観光地間の類似度測定に有効な情報である．本手法では，Yahoo 知恵袋とブログ記事を対象に，共起する単語を抽出し，その重要度を測定する．情報の取得には，Yahoo 知恵袋質問検索 API[*3] と Yahoo ブログ検索 API[*4] を利用する．Yahoo 知恵袋（以降，知恵袋）には，質問ごとに「エンターテインメントと趣味」や「暮らしと生活ガイド」のようなカテゴリ情報が付与されている．その中で観光と密接な関係を持つと考えられる「地域，旅行，お出かけ」カテゴリ（以降，地域カテゴリと呼ぶ）の質問と回答を処理の対象とする．また，知恵袋では，各質問に対して最良の回答をベストアンサーと定義している．この知恵袋の特性に注目し，任意の観光地を検索クエリとし，その際にヒットした質問のベストアンサーのみを共起度算出のために利用する．一方，ブログに関しては，同様に任意の観光地を検索クエリとし，それにヒットしたブログタイトル・サマリーを使用する．

キーワード抽出処理では，まず，得られた文章群について形態素解析を行い，単語に分割する．形態素解析器としては，MeCab[*5] を利用する．次に，得られた単語列からキーワードとなる候補を抽出する．この処理ではキーワード自動抽出ツールである TermExtract[*6] を使用する．TermExtract は単語の組み合わせを考慮し，出現頻度に基づいて重要度が算出される．

しかし，これらを用いて得られたキーワードには，例えば観光に直接関係ない語など，多くのノイズが含まれ

[*1] http://ja.wikipedia.org/wiki
[*2] http://chiebukuro.yahoo.co.jp
[*3] http://developer.yahoo.co.jp/webapi/chiebukuro/chiebukuro/v1/questionsearch.html
[*4] http://developer.yahoo.co.jp/webapi/search/blogsearch/v1/blogsearch.html
[*5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[*6] http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

表 1　各単語と S_k の値.

k	L_k	A_k	S_k
福岡	27647	164219	0.168
福岡タワー	212	325	0.652
明太子	450	4566	0.099
九州工業大学	2	711	0.003
東京	129547	886994	0.146
東京スカイツリー	2171	3984	0.545

表 2　「東京スカイツリー」の共起キーワードランキング.

ランク	知恵袋	ブログ
1	スカイツリー	スカイツリー
2	東京	東京
3	浅草	東京タワー
4	レジャー	東京ソラマチ
5	エンタメ	かめそば

る．そこで，事前に観光に関連した語を収集し，ノイズを除去する．ここでは，日本版 Wikipedia の見出し語約 100 万件を記述したテキストファイルを基に，観光に関連する語を収集する．まず最初に，英数字・記号だけで構成されている語は不要語とみなし，それらを削除する．さらに，残った単語一つ一つに対して，知恵袋の地域カテゴリ内で検索を行い，そのカテゴリ内のエントリでヒットする場合は，観光関連語として以降の処理の対象とする．今回，処理の対象となる語の数は 67644 語である．

次に，各単語の重要度を判定する．まず，ある単語 k について，地域カテゴリ内での総ヒット数を算出し，その総ヒット数を L_k とする．同様に，単語 k について，知恵袋の全カテゴリを対象に検索を行い，その総ヒット数 A_k も取得する．この 2 つのヒット数を基に，ある単語 k がどの程度観光に関連しているか (S_k) を

$$S_k = L_k / A_k \tag{1}$$

によって計算する．表 1 は，いくつかの語に対する L_k および A_k の値の例を示している．この表から，具体的な観光地の候補として考えられる「福岡タワー」や「東京スカイツリー」などの値は高くなり，観光と直接関係がないような語（ここでは大学名）は極めて低い値を持つことが分かる．

この値 S_k と各単語 k の出現頻度に基づき算出される値 F_k から

$$I_k = F_k \times S_k \tag{2}$$

によって，各単語の最終的な重要度 I_k を算出する．ここで，F_k は知恵袋もしくはブログ API で検索したときに得られる単語 k の頻度を C_k とした場合[*7]，

$$F_k = C_k^{\lambda} \tag{3}$$

で計算される．λ は単語頻度を減衰させる係数であり，$0 < \lambda < 1$ である．これは単純に単語の頻度をそのまま利用すると観光の関連度合いである S_k の影響力が弱まり，適切な重要度が得られないためである．λ の値をいくらとするかはキーワードのランキングに影響を与えるため，今回はいくつかの値を試したのち，経験的に $\lambda = \frac{1}{8}$ とした．

表 2 は，「東京スカイツリー」に関連するキーワードをソースごと（知恵袋とブログ）にランキングした結果である．この結果から，両方のソースで有効な共起キーワードが得られていることが分かる．

2.2　時系列分布

次に着目するのは時系列情報である．観光地には，その観光地が活性化する特定の時期やパターンが存在すると考えられる．ここでは，似たような活性化をする観光地は類似しているという仮定の基，この時系列情報を類似度測定のために特徴化する．

具体的には，2 つの時系列パターンを用いる．1 つは半年ごとの時系列パターンで，もう 1 つは半月ごと時系列パターンである．期間ごとに単語のヒット数を特徴とする．ここでも前節同様に知恵袋とブログを処理の対象とし，それぞれ最大で 1000 件の合計 2000 件のデータからヒット数を求める，

図 2 は，「清水寺（京都府）」と「天王寺公園（大阪府）」に関する半月単位の時系列パターンである．この図から，両方の観光地とも 10 月前後にヒット数が多いことが分かる．これは，共に紅葉シーズンに活性化する観光地であることを意味しており，このような時系列での類似性を考慮すれば，その観光地の特色を踏まえた類似性が判断できる．

2.3　カテゴリ情報

3 つめの特徴は，カテゴリ情報の利用である．2.1 節では，単語の観光らしさを測定するために，観光に関連していると考えられる「地域カテゴリ」のみに着目した．一方，本節では，2.1 節で対象とした「地域カテゴリ」以外のカテゴリに着目する．2.1 節で獲得された単語群は，他のカテゴリでも出現する．例えば，表 1 から東京スカイツリーは知恵袋で 3984 回出現したが，そのうち 1813 回は「地域カテゴリ」以外で出現していることが分かる．

[*7] 知恵袋を対象とした場合，C_k は「地域カテゴリ」での出現頻度であり，L_k と同じ値になる．

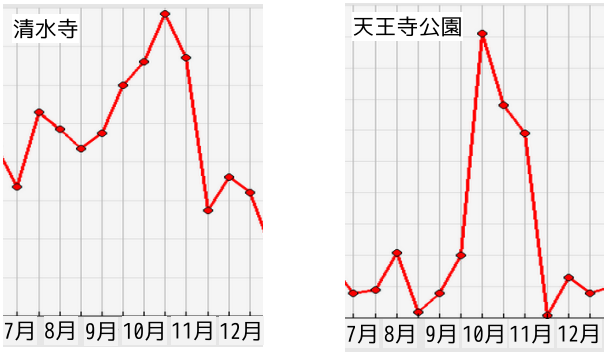


図 2　清水寺（左）と天王寺公園（右）の時系列データの一部

表 3　「東京タワー」のカテゴリ情報.

カテゴリ	件数
教養と学問，サイエンス	508
スポーツ，アウトドア，車	42
エンターテイメントと趣味	424
ニュース，政治，国際情勢	402
マナー，冠婚葬祭	24
暮らしと生活ガイド	114
健康，美容とファッション	17
生き方と恋愛，人間関係の悩み	144
ビジネス，経済とお金	37
コンピュータテクノロジー	3
子育てと学校	17
職業とキャリア	20
おしゃべり，雑談	2
インターネット，PC と家電	305

この「地域カテゴリ」以外でのどのカテゴリで出現するかという傾向は，その単語の持つ潜在的な特徴を表現していると考えられる．

表 3 は東京タワーに関するカテゴリ情報を表している．たとえば，ここから「教養と学問，サイエンス」のカテゴリの件数が多い事が分かる．これは建築学の視点で東京タワーの構造が話題になることが多いためである．このような，観光とは直接関係ない部分の情報を取り入れることで，新たな視点から（この場合は建築学的な視点や建物の構造が似ているものなどの視点から）類似した観光地を推薦できるようになる．

2.4　周辺情報

4 つめの特徴は，周辺施設に関する情報である．ここで周辺施設の情報とは，対象となる観光地の周辺 5km 以内にある施設のジャンル情報を指す．本論文において

表 4　「東京スカイツリー」の周辺情報.

周辺情報	件数
和食	4356
洋食	1356
中華	849
バイキング	25
ドラックストア	839
家電・携帯電話	976
コンビニ	1293
ディスカウントショップ	366
生活用品・インテリア	2526
趣味・スポーツ	642
ファッション・アクセサリー	2738

図 3　金閣寺（左）と福岡タワー（右）の標準マップ

は，Yahoo ローカルサーチ API[*8]を利用して得られる地図データから周辺情報を抽出する．

具体的には観光地の位置情報を入力として，58 種類の周辺情報を取得する．表 4 に東京スカイツリーに関する周辺情報の一部を示す．東京スカイツリーは都心部に位置しているため，コンビニや買い物施設，レストランなどの件数が多い．このような特徴から，各観光地がどのような場所にあるのかという地域情報やその傾向を推定することが可能になる．

2.5　地図画像

最後の特徴は，地図画像そのものから得られる特徴量である．2.4 節と同様に，対象となる観光地の緯度・経度から Yahoo スタティックマップ API[*9]を利用して地図画像を取得する．このとき，標準的な地図画像（以降，標準マップ）とミッドナイトモードの地図画像（以降，ナイトマップ）の 2 つを取得する．

[*8] http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/localsearch.html
[*9] http://developer.yahoo.co.jp/webapi/map/openlocalplatform/v1/static.html

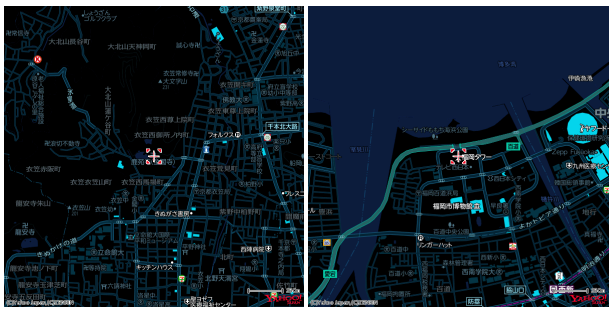


図 4　金閣寺（左）と福岡タワー（右）のナイトマップ

図 3 に標準マップの例を，図 4 に標準マップと同じ場所のナイトマップの例を示す．図から分かるように，標準マップでは海や川，山のような地理的情報が色の分布で獲得できる．この情報を特徴化することで，観光地周辺の地理的な情報（たとえば，海に面しているとか自然が豊かな場所にあるなど）を扱えるようになる．一方，ナイトマップでは，交通網や人工的な建物の量が可視化されている．この画像から，観光地周辺の都会度のような尺度が獲得できる．この画像の例では，金閣寺は図 3 において緑が多く含まれることから比較的自然があり，一方で図 4 で交通網も発達していることから適度に都市的な要素も含んでいる場所にあることがわかる．福岡タワーは図 4 から都会にあり，さらに図 3 から海に面していることがその画像中に含まれる色の分布で推測できる．これらの情報は 2.4 節の周辺情報だけでは獲得できない特徴である．

提案手法では，これらの画像から色情報を抽出する．色情報の抽出には PHP のクラスライブラリ ColorExtract を利用する．今回は HTML で表現されるカラーコードを対象とする．抽出された各色の画像中での出現頻度を各画像の特徴量として利用する．

3　推薦システム

本節では，本論文で提案する推薦システムについて説明する．まず，基盤となる観光地データベースの構築手法について説明し，次に推薦に利用する類似度計算の枠組みについて説明する．最後に推薦システムで利用可能な 2 つのパターンの推薦方法について説明する．

3.1　観光地データベース

本システムを実現するにあたって，観光地の情報を収集した観光地データベースが必要となる．そこで，2.1 節で作成・活用した観光キーワード辞書から観光地を抽出することで観光地データベースを構築する．

この観光地抽出処理では，2.4 節でも述べたローカルサーチ API を活用する．ローカルサーチ API では，任意のキーワードを入力とした場合，そのキーワード名と関連性の高い施設情報を取得することができる．これを利用することで，観光地の抽出を実現する．また，その際に位置情報（緯度経度・都道府県名）も取り出し，これらを観光地データベースに登録する．これにより，4507 件の観光地を登録した観光地データベースの構築が完了する．

最後に，観光地データベースに登録されている観光地情報を対象に，2 節に基づき特徴情報をすべて収集する．その結果を観光地データベースに組み込むことで，各観光地についての情報が特徴量化された観光地データベースが作成される．

3.2　類似度測定

提案システムでは，2 節で述べた観光地の特徴量から特徴ベクトルを作成する．本節では，そのベクトルを用いた類似度計算処理について述べる．以下，対象とする観光地を s で表現する．

共起キーワードリストの特徴ベクトル（以下，共起キーワードベクトルと呼ぶ）への適用について述べる．共起キーワードベクトル Key_s は，

$$Key_s = \{x_1, x_2, \ldots, x_{n_k}\} \tag{4}$$

と表現できる．ここで，n_k は観光キーワードリストの登録されているキーワード数 67644 語となる．共起キーワードベクトル Key_s 内の各要素 (x) は，各共起キーワードの重要度の値が入る．

時系列データの特徴ベクトル（以下，時系列ベクトルと呼ぶ）への適用について述べる．時系列ベクトル $Time_s$ は，

$$Time_s = \{x_1, x_2, \ldots, x_{n_t}\} \tag{5}$$

と表現できる．ここで，n_t は，時系列データの特徴パターンである 2004 年前期～2012 年後期の計 18 通りと，1 月前半～12 月後半の計 24 通りの合計 42 通りとなる．時系列ベクトル $Time_s$ 内の各要素は，各時期でのヒット数の値が入る．

カテゴリ情報の特徴ベクトル（以下，カテゴリベクトルと呼ぶ）への適用について述べる．カテゴリベクトル $Cate_s$ は，

$$Cate_s = \{x_1, x_2, \ldots, x_{n_c}\} \tag{6}$$

と表現できる．ここで，n_c は，知恵袋上での地域カテゴリを除く最上位のカテゴリ数 14 となる．カテゴリベクトル $Cate_s$ 内の各要素は，各カテゴリ内でのヒット数の値が入る．

周辺施設情報の特徴ベクトル（以下，周辺施設ベクトルと呼ぶ）への適用について述べる．周辺施設ベクトル

$Surr_s$ は，

$$Surr_s = \{x_1, x_2, \ldots, x_{n_e}\} \tag{7}$$

と表現できる．ここで，n_e は，施設情報の第二階層ジャンル数 58 となる．周辺施設ベクトル $Surr_s$ 内の各要素は，観光地 s の周辺 5km 以内で検索した時のヒット数の値が入る．

地図画像データの特徴ベクトル（以下，マップベクトルと呼ぶ）への適用について述べる．マップベクトル Map_s は，

$$Map_s = \{x_1, x_2, \ldots, x_{n_m}\} \tag{8}$$

と表現できる．ここで，n_m は，カラーコードのパターン数である 16^6 となる．マップベクトル Map_s 内の各要素は，カラーコードの使用回数の値が入る．

提案システムでは，類似度計算処理としてコサイン類似度を適用する．上記の 5 つのベクトルを統合し，観光地ごとに $Vec_s = \{Key_s, Time_s, Cate_s, Surr_s, Map_s\}$ を構築する．このとき，ある観光地 $s1$ と $s2$ の類似度は以下の式で計算される．

$$Cos(s1, s2) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \times \sum_{i=1}^{n} y_i^2}} \tag{9}$$

ここで，x_i と y_i は観光地 $s1$ と $s2$ に対する特徴ベクトル Vec_{s1} と Vec_{s2} の各要素である．

3.3 推薦方法

提案システムでは，前述のように観光地間の類似度を測定し，その値に基づいて観光地を推薦する．そのため，ユーザからシステムへ推薦の基準となる観光地の入力が必要になる．図 5 は構築した提案システムの入力画面である[*10]．提案システムに対象となる都道府県を入力すると，観光地データベースからその都道府県に存在する観光地のリストが検索され，提示される．ユーザは，そのリストの中から対象となる観光地を選択するか，入力フォームに文字列を入力して観光地を検索する．

提案システムでは，お薦め観光地の提示に関して，ユーザの置かれている環境に応じて 2 つの戦略を用意している．ユーザは以下に示す 2 つの想定環境に応じて，2 つの戦略のうち 1 つを選択する．

1 つめの想定環境は，あるユーザが旅先で次の観光地を探している場面である．すなわち，今いる観光地が上記の入力となり，それがユーザの嗜好を反映しているという仮定を置くことができる．この場合，ユーザは図 5 の「（3）提案観光地と入力観光地の「距離範囲」」という部分に検索の対象とする距離範囲を指定する．現在は，

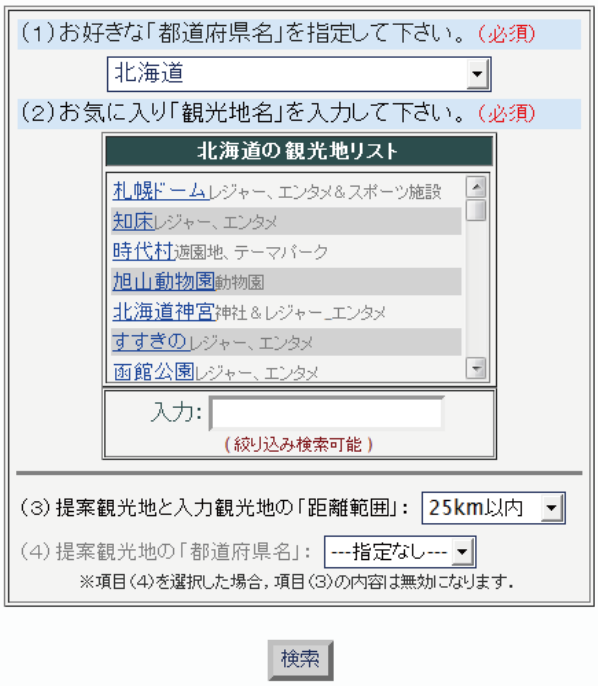


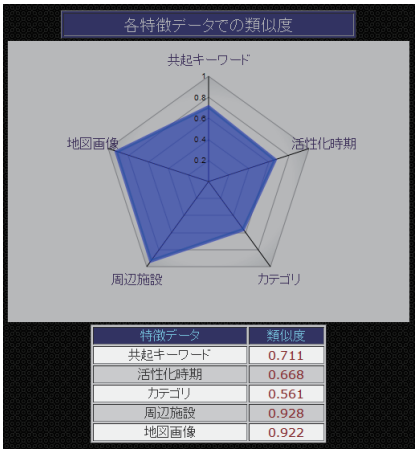
図 5　提案システムの入力画面．

5km 以内，10km 以内，25km 以内，50km 以内，100km 以内の 5 つの距離範囲を指定できる．システムは，画面中の「（2）お気に入り「観光地名」を入力して下さい。」の部分で入力された観光地とそこを基点としてこの距離範囲内に存在する観光地に対して類似度を計算し，それを類似度に基づきソートして提示する．

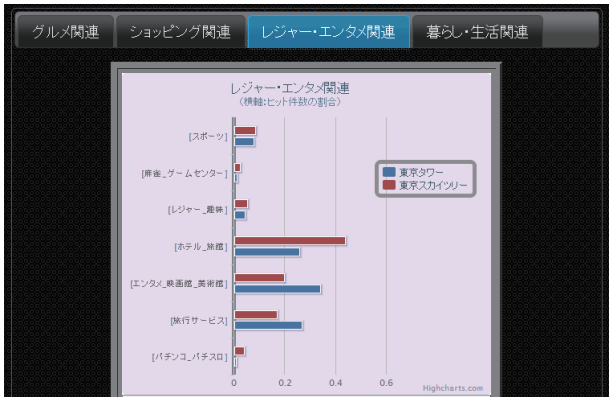
2 つめの想定環境は，特定の都道府県への旅行計画などを事前に立てる場合の利用である．すなわち，入力は自分の好きな観光地であり，出力は旅行予定の都道府県でユーザの好みそうな観光地となる．この場合，ユーザは図 5 の「（4）提案観光地の「都道府県名」」に旅行を計画している都道府県名を入力する．システムは，ユーザの好きな観光地を基に，指定された都道府県内に存在する観光地との類似度計算を行い，その類似度に基づいて推薦観光地を提示する．

それぞれの戦略に基づいて提示されたソート済みのリストからユーザが気になる観光地をクリックすることで，入力した「ユーザの嗜好を反映した観光地」と「推薦された観光地」の類似性が，2 節で示された 5 つの特徴毎にレーダチャートやグラフによって可視化され，比較が可能になる．図 6 は「東京タワー」と「東京スカイツリー」の総合的な類似性と周辺施設の一部について比較したグラフの例である．これらの情報から，この 2 つの観光地が特に周辺環境や地理的な面でよく似ていることが分かる．

[*10] 提案システムは Web 上で動作している．
http://tlr.pluto.ai.kyutech.ac.jp/



すべての類似度比較画面



周辺施設に関する比較画面の一部

図 6　提案システムの出力例.

4　動作検証・考察・議論

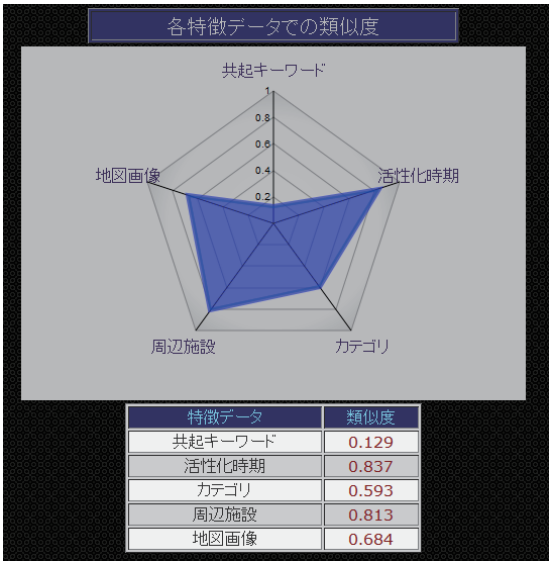
本節では，実装された観光地推薦システムの動作例を基に，その有効性について検証する．また，集合知の観点から，今後の拡張点について考察および議論する．

一般的な観光地推薦システムや観光やイベントに関連する情報の抽出処理では，その観光地やイベントに関するテキストなどを分析し，その内容に基づいて類似性を判断することが多い．安村ら [4] は Blog を対象として，イベントの発生時間や場所などの情報を抽出する手法を提案している．倉島ら [5] も Blog を対象として，街の話題を抽出し，地図上に提示するシステムを構築している．提案手法では，2.1 節で述べた共起キーワードのみを用いた手法などがそれらと類似する．しかし，有名ではない観光地では，その観光地に関する記述量が十分ではなく，入力された観光地との類似性を適切に判断できないことも多くある．また，Blog などのテキストを対象とした場合，ノイズとなる単語なども多く，正しく類似性を判断できない場合も多い．一方で，提案するシステムでは，キーワード以外の観点からも類似度を測定できるため，十分なテキストが存在しないような観光地についても，少なくとも地図画像などから類似性が判定可能になる．このような点から，複数の情報を統合的に利用する提案システムの有効性を見ることができる．

活性化時期という時系列データの類似性を判断することで，テキストでは言及されにくいもしくは検出が難しい，季節的な観光地の類似性を測ることが可能になる点も提案システムの有用な部分である．図 7 は福岡の「太宰府天満宮」を入力として，東京を対象にした場合に上位に出力される「湯島天神」の全体の類似性と活性化時期に関するグラフである．これら 2 つはともに学問の神様として有名であり，それが活性化時期として共に顕著化することで高い類似性を持つようになる．初詣シーズンの 1 月以外にも，グラフから 2 月〜3 月頃の受験シーズンに活性化することがわかる．一方で，すべての類似性のグラフから分かるように，共起キーワードだけでは一致する文字列が十分ではなく，この 2 つの類似性を正確に捕捉できない．このような点も，多角的に観光地の類似性を判断できる提案システムの有効性を示している．

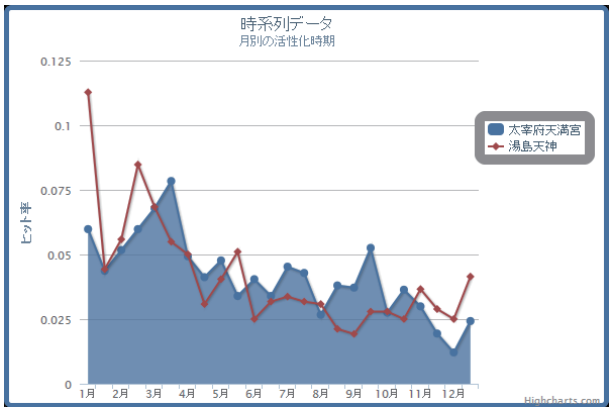
また，地図画像を利用する点も大きな利点である．図 8 は「福岡タワー」を基点とし，神奈川県を対象とした場合に上位に出力される「パシフィコ横浜」の地図画像と色分布を表している．地図画像を利用することで，海に面しているという地理的な条件を加味した推薦が可能になっている．そのほかの事例として，図 9 のようなものも存在する．これは東京にある「亀戸天神社」を入力とし，福岡県を対象に検索した結果の一部である．「亀戸天神社」が菅原道真を祀っているという観点から考えれば，その点で際立って有名な前述の「太宰府天満宮」が上位で推薦されると考えるのが直感的である．しかし，この検索結果では，「太宰府天満宮」は 15 位に位置し，一方で 3 位に同じく菅原道真に関連する「水鏡天満宮」がランクインした．これは，入力である「亀戸天神社」の地域的特性（都会度）が地図画像により評価され，「太宰府天満宮」よりも都市部に存在する「水鏡天満宮」が優先された結果である．この結果，すなわち「水鏡天満宮」が「太宰府天満宮」よりも優先されるということが妥当かどうかは人によって異なるため，議論の余地が残るが，提案手法が様々な観点から様々な観光地を推薦できていることの実例であることは確かである．このような推薦をテキストだけで行うことは困難である．

以上に示したように，複数の情報源や種類の異なる情報（たとえば，テキストと画像）を統合的に扱う提案手法は，1 つの情報源に依存するシステムよりも多くの点で優れていることが分かる．今回は Yahoo 知恵袋や



特徴データ	類似度
共起キーワード	0.129
活性化時期	0.837
カテゴリ	0.593
周辺施設	0.813
地図画像	0.684

太宰府天満宮と湯島天神のすべての類似度



太宰府天満宮と湯島天神の活性化時期のグラフ

図 7　時系列データの有効性.

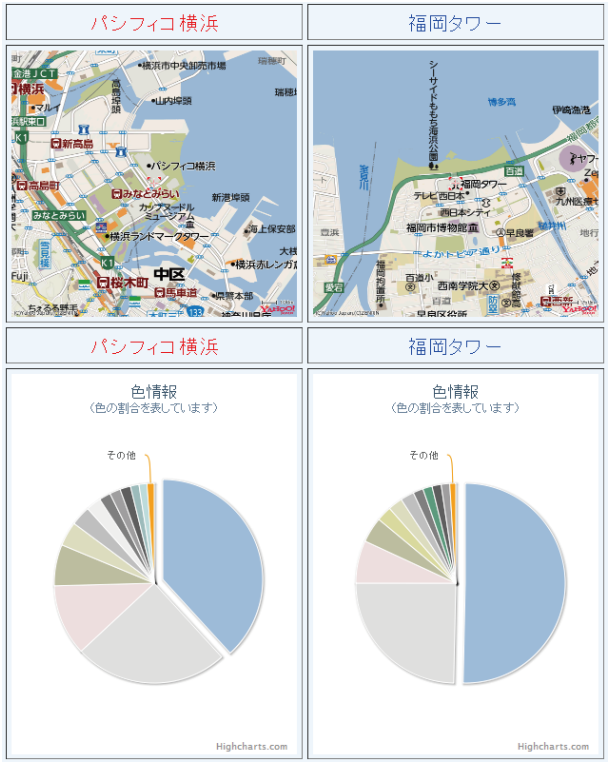


図 8　地図画像の有効性（地理的情報）.

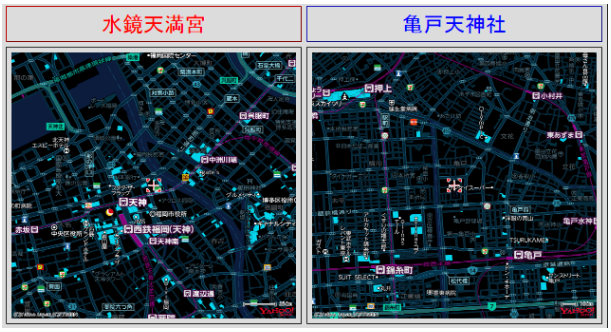


図 9　地図画像の有効性（都会度）.

Blog，Wikipedia，地図データなどを利用してシステムを構築したが，Web 上にはそれ以外にも多くの情報源が存在する．我々は，Twitter を対象とした観光情報の分析 [6] や Tweet が観光地などで実際につぶやかれているかどうかの判定 [7] に関する研究などを行っている．これらの研究の結果との統合や分析も今後の課題である．

長尾 [8] は写真共有サイトである Panoramio を利用した観光情報の提供方法について考察している．その論文では，写真が密集している地域は訪問者の多い観光地である可能性が高いことを示している．これを我々の研究に転用すれば，写真が頻繁にアップロードされる時期は，活性化時期の特徴量を別の情報源から捉えることと等価になる．実際に，Panoramio のデータを分析した結果，知恵袋から獲得した時系列データと Panoramio から獲得した時系列データには高い類似性が見られた．一方で，この 2 つが違う傾向を見せた例を図 10 に示す．この例では，1 月に Panoramio のデータが突出しいる．これは，この 2 つのデータが持っている特性の違いである．具体的には，Yahoo 知恵袋はその場にいる人，行きたい人，行った人などの様々な投稿者がいるが，Panoramio は「写真を撮影して投稿する」というサービス自体の持つ特性からその場にいる人のみが対象になっているためである．実際に図 10 の例で，1 月に投稿された写真を見てみると，「初日の出」を撮影し，投稿したものが多かった．このように，Web 上の別の情報源を利用することは，推薦のために有効な新しい特徴量の獲得に繋がる．これらの情報を現在公開しているシステムに組み込むことは重要な課題である．

提案システムでは，3.3 節で述べた 2 つの想定環境に基づき，ユーザからの入力を具体的な 1 つの観光地とその探索範囲（距離か場所か）として設計した．したがって，複数の観光地や「綺麗な」や「緑豊かな」などの抽象的表現の利用など，ユーザからどのような入力を受け

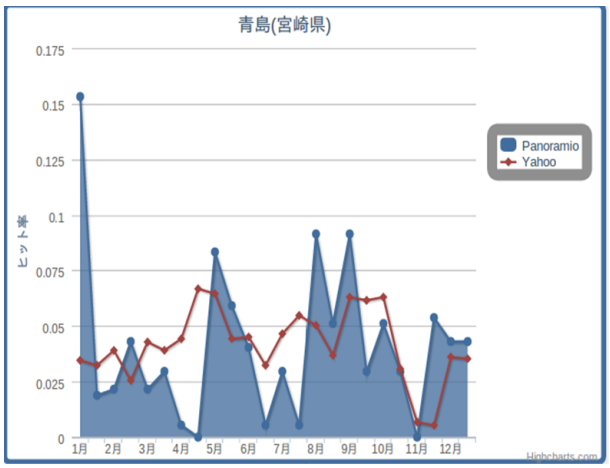


図 10　Yahoo 知恵袋と Panoramio の活性化時期の比較.

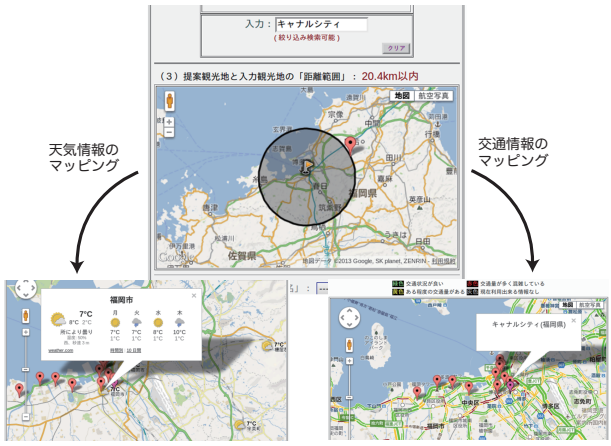


図 11　地図上への結果の投影.

取ることが最適かという入力の多様性については十分に議論されていない．これは，本論文の目的が複数の情報源を利用して類似度を算出すること，すなわち情報源の多様化とその有効性検証であり，入力の多様性については，別の問題として捉えているためでもある．しかしながら入力の多様化は重要な課題であり，システムの実証実験などに基づき，今後考察していく必要があるだろう．

出力に関しては，現在は類似度に基づき，観光地をリストの形式でユーザに提示しているが，図 11 に示すように，マップ上に提示することも検討しており，実装・評価中である．リストではなく地図として可視化することで，その場所の様々な情報（天気や交通情報など）と統合して提示できるため，ユーザにとって，より有益な情報を含み，なおかつ可読性が高い出力が可能になる．また地図を利用することで，奥山らの研究 [9] のように観光経路の推薦なども可能になると考えられる．この点についても今後十分な考察が必要である．

5　おわりに

本論文では，集合知に基づく観光地推薦システムを提案し，実装した．情報源として，Yahoo 知恵袋や Blog，地図情報などを利用した．作成したシステムの動作例を基にシステムの有効性を検証した．

今後の課題としては，Twitter などのテキスト情報の有効利用とシステムへの統合や Panoramio に代表されるような画像データの利用と統合，さらに推薦システムのインタフェースの改良による，入力の多様性やデータの可読性向上などが挙げられる．

参考文献

[1] 川村秀憲, 鈴木恵二, 山本雅人, 松原 仁：観光情報学, 情報処理学会デジタルプラクティス, Vol. 3, No. 4, 2012.

[2] 斎藤 一：Web における観光情報提供と分析, 人工知能学会誌, Vol. 26, No. 3, pp. 234-239, 2011.

[3] Yuya Kanazawa, Yosuke Hidaka and Katsuhiko Ogawa: Destination Retrieval System using an Association Retrieval Method, International Journal of Future Computer and Communication, Vol. 2, No. 3, pp.169-173, 2013.

[4] 安村祥子，池崎正和，渡邉豊英，牛尼剛聡：blog マッピングを用いたイベント情報抽出，第 18 回データ工学ワークショップ (DEWS2007)，2007.

[5] 倉島健，手塚太郎，田中克己：Blog からの街の話題抽出手法の提案，電子情報通信学会第 16 回データ工学ワークショップ (DEWS2005)，2005.

[6] Kazutaka Shimada, Shunsuke Inoue, Hiroshi Maeda and Tsutomu Endo: Analyzing Tourism Information on Twitter for a Local City, Proceedings of SSNE2011, International Workshop on Innovative Tourism, 2011.

[7] Kazutaka Shimada, Shunsuke Inoue and Tsutomu Endo: On-site likelihood identification of tweets for tourism information analysis, Proceedings of 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM 2012).

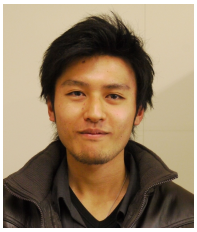
[8] 長尾光悦：CGM をベースとした観光情報提供方法に関する考察，観光情報学会 第 9 回全国大会，pp. 26-27，2012.

[9] 奥山幸也, 柳井啓司：写真撮影の位置軌跡を利用した旅行支援システム， DEIM Forum，F7-6，2011.

著者紹介

嶋田 和孝（会員）昭和 49 年生．平成 9 年大分大学工学部知能情報システム工学科卒業．平成 11 年同大大学院博士前期課程修了．平成 14 年同博士後期課程単位取得満期退学．同年より九州工業大学情報工学部助手．平成 24 年より大学院情報工学研究院知能情報工学研究系准教授．博士（工学）．評判分析をはじめとする Web 上のテキスト分析や対話の理解に関する研究に従事．観光情報学会，言語処理学会，電子情報通信学会，情報処理学会，人工知能学会各会員．

上原 尚（非会員）平成 2 年生．平成 25 年九州工業大学情報工学部知能情報工学科卒業．現在，(株)NEC 情報システムズに勤務．在学中は観光情報の分析，マイニングに関する研究に従事．



遠藤 勉（非会員）昭和 24 年生．昭和 47 年九州大学工学部電子工学科卒業．昭和 49 年同大大学院修士課程修了．昭和 52 年同博士課程単位取得退学．同年より九州大学工学部助手，昭和 55 年に大分大学工学部講師，昭和 56 年に同大学助教授，平成 4 年同大学教授を経て，平成 12 年から九州工業大学情報工学部教授．現在は九州工業大学名誉教授．工学博士．自然言語処理，コンピュータビジョンの研究に従事．