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

寺西拓也[1]　野村達二[1]　平山智子[1]　石野亜耶[2]　難波英嗣[2]　竹澤寿幸[2]

[1]広島市立大学 情報科学部　[2]広島市立大学大学院情報科学研究科

1. はじめに

観光は地域における消費増加や雇用の創出など幅広い経済効果をもたらすことから、観光立国の実現を目標に掲げ、2008年10月に観光庁が設置され、観光を21世紀における日本の重要な政策として位置づけた多様な取り組みが推進されている。観光を支援する媒体としては、旅行会社などが運営する観光ポータルサイトや、旅行情報雑誌「るるぶ」などの観光情報データベースなどが作成、公開されている。旅行をする際には、このような旅行情報雑誌は、初めて訪れる都市に関する情報を収集するための手段として必要不可欠である。旅行情報雑誌「るるぶ」のような観光ガイドブックには、観光地の人気スポットなど、様々な情報が記載されている。しかし、眺めた景色や食べ物などに関する感想などの人によって意見や感想が異なる情報は多くの意見を必要とし、観光ガイドブックの情報だけでは十分でない。そこで、多様な観光情報の情報源として SNS (Social Networking Service)に着目した。そこで本研究では、観光ガイドブックの各ページを「見る」「体験する」「買う」「食べる」「泊まる」のカテゴリに分類することによる構造化と、観光ガイドブックへ旅行ブログエントリと質問応答コンテンツの対応付けを行う。例えば、「見る」のカテゴリページには、各ガイドブックに対応付けされている旅行ブログエントリと質問応答コンテンツを対応付ける。

本論文の構成は以下の通りである。2節では関連研究、3節でシステム構成と提案手法、4節では実験結果について述べ、5節で本稿をまとめる。

2. 関連研究

本研究と同様にテキストに情報を対応付けする研究としてRakeshら[1]の研究がある。Rakeshらは、文字が多く視覚的な資料が不足している発展途上国の教科書に関連する画像を対応付けしてすることにおり、テキストを分かり易くしている。また、Rakesh らは教科書に関連する画像をWEBサイトから検索する手法を提案している。

また、石野ら[2]は、旅行ブログから観光情報を自動的に収集する研究を行っている。石野らは、旅行ブログエントリから自動的に観光情報リンク(旅行ブログエントリ中に含まれるリンク)を収集し分類する手法を提案している。さらに、観光情報リンクに対し関連性の高い楽天市場の商品を発見し、その商品への宣伝リンクを自動的に付与する手法を提案している。

柴崎ら[3]はYahoo! 知恵袋の「地域」のカテゴリを用いた研究を行っている。柴崎らはQ&Aサイトに投稿される「地域」に関する質問及び回答は対象地域に対するユーザの要求・関心をより直接的に反映していると考えた。そこで、Yahoo! 知恵袋を対象に東日本大震災被害地域である岩手県に関する質問とYahoo! JapanサイトのWebニュースを収集し、Yahoo! 知恵袋において表出したユーザの地域に対する要求や関心がどのようにユーザ間で共有されているのか分析するのと同時に、ユーザの要望・興味を時系列的な推移を可視化するシステムを実装している。

本研究と同様に、SVM を用いて分類する研究として劉ら[4]の研究がある。劉らも観光に関する研究で SVM による分類を行っている。劉らはコーパスに基づく観光案内システムの構築を行っている。

3. 観光ガイドブックと観光コンテンツの対応付け

3.1節ではシステム構成、3.2節では観光ガイドガイドブックの構造化、3.3 節で対応付けについて述べる。

3.1 システム構成

本節では、最終的に目指す観光ガイドブックとSNS 上の観光情報などを観光ガイドブックと自動的に対応付けするシステム構成について述べる。

まず、観光ガイドブックを構造解析し、カテゴリ分類を行う。次に、観光ガイドブックに観光コンテンツを対応付けする。そして、対応付けした観光コンテンツを観光ガイドブックと同様にカテゴリ分類し、各ガイドブックのカテゴリごとに観光コンテンツを対応付けする。

3.2 観光ガイドブックの構造化

- **使用する観光ガイドブック**

本研究では JTB パブリッシングが発行している観光ガイドブック「るるぶ」と昭文社が発行している「まっぷる」をスキャンし、OCR によりテキスト化したデータを使用する。

- **カテゴリの種類**

カテゴリを表1に定義する。

表 1：カテゴリの判定基準

カテゴリ	判定基準
見る	観光名所のページなどの自分は動かなくても見るだけで楽しめる物について書かれたページ。
体験する	○○体験やスキューバダイビングなどの自分の体を使って楽しめる物について書かれたページ。
買う	お土産に関するページ。
食べる	飲食店の情報が載っているページ。
泊まる	ホテルの情報が載っているページ。
その他	「見る」、「体験する」、「買う」、「食べる」、「泊まる」に該当しないページ。例として広告ページや巻末の交通情報のページ。

- **観光ガイドブックのカテゴリ分類**

本研究では、人手で収集した手掛かり語を機械学習の素性として用い、観光ガイドブックの各ページを表 1 に示すカテゴリに自動分類する。人手で収集した手掛かり語の例を表 2 に示す。

表 2　手掛かり語の例

カテゴリ	手掛かり語	単語数
見る	世界遺産、景色、博物館	48 語
体験する	トレッキング、カヌー	144 語
買う	おみやげ、免税店、銘菓	165 語
食べる	レストラン、ランチ、名物	36 語
泊まる	泊まる、リゾート、客室	22 語

3.3 観光ガイドブックと観光コンテンツの対応付け

観光ガイドブックでは得られない情報を得る手段として、我々の研究では、「旅行ブログエントリ」と「Yahoo! 知恵袋」に着目し、これらのコンテンツを観光ガイドブックに対応付けする。ガイドブックとこれらのコンテンツを対応付ける際、「地名」は大きな手掛かり語となると考えた。

本研究では、まず、日本語構文解析器 CaboCha を用いて地域名(LOCATION)と組織名(ORGANIZATION)を抽出し、これらを内容語として用いて索引語リストを作成する。次に、以下の 8 つの手法により、ガイドブックへの旅行ブログエントリおよび Yahoo! 知恵袋の対応付けを行う。なお、対応付けには、汎用連想計算エンジン GETA を用いる。4 分割交差検定により、訓練用データで F 値が最も高くなる値を閾値に決め、評価用データで類似度が閾値以上のものをシステムが正解と判定したデータとする。

旅行ブログエントリの 4 手法

◆LOC: 「地名」のみの頻度を利用。
◆LOC+TITLE :「地名」のみの頻度を利用（ただし旅行ブログエントリのタイトルとなっている地名については頻度を 10 とする)。
◆LOC+ORG: 「地名」と「組織名」の頻度を利用。
◆LOC+ORG+TITLE : 「地名」と「組織名」の頻度を利用。(ただし旅行ブログエントリのタイトルが地名または組織名については頻度を 10 とする)

Yahoo! 知恵袋の 4 手法

◆LOC(Q)：知恵袋の質問文中の「地名」のみの頻度を利用。
◆LOC(Q&A)：知恵袋の質問文と回答文中の「地名」のみの頻度を利用。
◆LOC+ORG(Q)：知恵袋の質問文中の「地名」と「組織名」の頻度を利用。
◆LOC+ORG(Q&A)：知恵袋の質問文と回答文中の地名」と「組織名」の頻度を利用。

比較のため、ガイドブック中の単語（名詞、動詞、形容詞）を内容語として用いた手法を BASELINE1、知恵袋の質問文中の単語を内容語として用いた手法を BASELINE2、知恵袋の質問と回答文の単語を内容語として用いた手法を BASELINE3 として実験を行う。

上述の手法に加え、さらに精度を重視した以下の 2 手法でも実験を行う。
◆TOP1:類似度が最も高いガイドブックを出力。
◆THRE：類似度が最も高く、かつ類似度が閾値以上のガイドブックを出力。

4. 実験

3.2 節および 3.3 節で述べた提案手法の有効性を確認するため、実験を行った。3.2 節に関する実験を「観光ガイドブック構造解析実験」、3.3 節に関する実験を「SNS 対応付け実験」とする。

4.1「観光ガイドブック構造解析実験」

■　実験に用いるデータ

「観光ガイドブック構造解析実験」では、OCR によりテキスト化した観光ガイドブック 20 冊分(国内 10 冊、海外 10 冊)である 2,897 ページを実験対象とした。そして、人手により構造分類された結果と、システムにより判定した結果の比較を行う。人手により判定された結果を表 3 に示す。

表 3：実験に用いる正解データの内訳

カテゴリ	分類数(ページ)
見る	1,026
体験する	78
買う	418
食べる	741
泊まる	278
その他	356

■ 機械学習

カテゴリの分類には機械学習には、TinySVM を用いた。2 次の多項式カーネルを使用し、2 分割交差検定を行った。

■ 評価尺度

評価尺度は精度と再現率を用いた。

4.2「SNS 対応付け実験」

■ 実験に用いるデータ

「SNS 対応付け実験」では、観光ガイドブック 90 冊に対し、旅行ブログエントリと Yahoo! 知恵袋の対応付けを行う。

ブログエントリの対応付け実験では、石野ら[2]の研究により収集された旅行ブログエントリのうち、人手で正解判定を行った 489 件を使用した。

知恵袋の対応付け実験では、Yahoo! 知恵袋データの「地域、旅行、お出かけ」カテゴリのうち、人手で正解判定を行った 1,990 件を使用した。

■ 評価尺度

評価尺度は精度と再現率と F 値を用いた。

4.3 実験結果

「観光ガイドブック構造解析実験」の実験結果を表 4 に示す。

表 4：カテゴリ分類の実験結果

カテゴリ	精度(%)	再現率(%)
見る	69.02	34.61
体験する	84.21	25.05
買う	70.73	9.29
食べる	77.40	36.84
泊まる	69.70	26.23
平均	74.21	26.40

上記の実験結果より、各カテゴリにおいて精度に関して高い数値を記録した。よって提案手法の有効性を示せたと考えられる。

「SNS 対応付け実験」のブログエントリに関しての実験結果を表 5 に、知恵袋に関しての実験結果を表 6 に示す。

表 5：ブログエントリの対応付け実験結果

	精度(%)	再現率(%)	F 値(%)
BASELINE1	12.71	16.47	14.05
LOC	54.41	44.19	48.71
LOC+TITLE	56.57	42.76	48.43
LOC+ORG	53.78	45.66	48.95
LOC+ORG+TITLE	56.34	42.94	48.16

表 6：知恵袋の対応付けの実験結果

	精度(%)	再現率(%)	F 値(%)
BASELINE2	24.85	24.66	24.58
BASELINE3	24.67	24.29	24.41
LOC(Q)	65.26	28.86	38.57
LOC(Q&A)	50.46	38.42	42.33
LOC+ORG(Q)	45.18	45.66	44.51
LOC+ORG(Q&A)	57.02	35.88	43.33

表 5、6 より BASELINE 手法と比べ、どの手法でも F 値が向上しているため、「地名」、「組織名」の類似度で分類する手法の有効性が確認できた。

表 5 より最も精度の高い LOC+TITLE、表 6 より最も精度の高い LOC(Q)での類似度結果を用いて、ブログエントリでの実験結果を表 7、Yahoo! 知恵袋での実験結果を表 8 に示す。

表 7：精度重視ブログエントリの対応付け

	精度	再現率	F 値
TOP1	71.69	30.64	42.93
THRE	77.37	22.67	35.06

表 8：精度重視の知恵袋の対応付け

	精度	再現率	F 値
TOP1	85.18	16.72	27.96
THRE	92.40	8.13	14.94

4.4 考察

本節では、以下の 2 点に関して、実験結果の考察を行う。

① 「観光ガイドブック構造解析実験」の考察
② 「SNS 対応付け実験」の考察

① 「観光ガイドブック構造解析実験」の考察

全体的に再現率が低かった原因として、OCRで処理する際に文字を認識できなかったページや文字化けばかりのページがすることが考えられる。

カテゴリ分類の判定誤りの原因として、人手により判定する際には、観光ガイドブックに記載されている画像も参考にしていることが挙げられる。しかし、それらのページには、「食べる」の手掛かり語である「レストラン」や「泊まる」の手掛かり語である「ホテル」などの他のカテゴリの手掛かり語を含むこともあるため判定誤りが発生したと考えられる。この問題を解決する方法として、画像解析を用いてカテゴリ分類を行うことにより、精度・再現率の更なる向上が期待できる。

② 「SNS 対応付け実験」の考察

精度重視で行った対応付けの判定誤り（ブログエントリでは 40 件、知恵袋では 20 件）を種類ごとに分析し、原因を以下に示す。

◆要因 1　表記揺れ（ブログエントリ：0.0%、知恵袋：10.0%）

知恵袋の判定誤りの 20 件中 2 件は表記揺れが原因であった。旅行コンテンツから抽出された「地名」の中に「関空」いう名詞があった。しかし、このコンテンツデータを対応付けすべきガイドブックには「関西空港」という表記で記載されており、「関空」という略語が出現した、別の観光ガイドブックに対応付けされていた。

◆要因 2　観光ガイドブックの特徴（ブログエントリ：52.5%、知恵袋：55.0%）

ブログエントリにおいて 40 件中 21 件、知恵袋において 20 件中 11 件がガイドブックの特徴が関連する判定誤りであった。

公共交通機関を使用する旅行者は多い。そのため観光ガイドブックには、高速バスや新幹線、飛行機など交通手段について記載されたページが多く存在する。そのようなページには、隣の都道府県や他の大都市の駅名、空港名など、その観光ガイドブックの都市には存在しない地名も多く存在するため、判定誤りが発生したと考える。

◆要因 3　固有表現の重みが同じ（ブログエントリ：42.5%、知恵袋：30.0%）

ブログエントリにおいて 40 件中 17 件、知恵袋において 20 件中 6 件が観光コンテンツ内の固有表現の重みによる判定誤りであった。

観光コンテンツの本文中に 2 つ以上地名または組織名が存在した場合、重み付けを行っていないため誤って対応付けしてしまったと考える。ブログエントリの分類に関しては、TITLE に地名が存在した場合、重み付けを行っているため、TITLE に地名または組織名が存在した場合、「＋TITLE」付きの手法では改善されている。

◆要因 4　複数箇所に存在する固有表現（ブログエントリ：5.0%、知恵袋：5.0%）

ブログエントリにおいて 40 件中 2 件、知恵袋において 20 件中 1 件がガイドブックの特徴が関連する判定誤りであった。

「府中」や「日本橋」など複数箇所に存在する地名が抽出されたため、判定誤りが発生した。

5. おわりに

本研究では、観光ガイドブックの各ページをカテゴリ分類する手法と、観光ガイドブックと観光コンテンツを対応付けする手法を提案した。実験の結果、2 つの手法は共に高い精度でリンクタイプの判定を行うことができ、提案手法の有効性が確認された。

参考文献

[1] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Kishnaram Kenthapadi, “Enriching Textbooks With Images” 20th ACM Conference on Information and Knowledge Management (2011)

[2] 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)

[3] 柴崎真理子, 藤田秀之, 木實新一, 有川正俊, “Q&A サイトにおけるユーザの要求・関心の時空間的な推移の可視化”, 第 4 回知識共有コミュニティワークショップ論文集, pp.41-44 (2011)

[4] 劉曄, 滕智, 土屋誠二, 任福継, “VSM に基づく SVM と構造解析手法を用いた旅行案内システムの構築”, 情報処理学会研究報告.自然言語処理研究報告 (2008)