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

石野亜耶 [1,a)]　藤井一輝 [2]　藤原泰士 [2]　前田剛 [2]
難波英嗣 [1]　竹澤寿幸 [1]

概要：観光を支援する媒体として旅行ガイドブックが挙げられる．しかし，旅行ガイドブックに掲載されている情報は，一般的な情報であり，さまざまな年齢層や性別の旅行者が求める多様な情報は掲載されていないといった問題点がある．不足する観光情報を補うための情報源として，旅行者が旅行記を記述した旅行ブログエントリや，旅行に関連する疑問を質問する場となる質問応答コンテンツが挙げられる．そこで，本研究では，これらのコンテンツを旅行ガイドブックに自動的に付与することで，旅行ガイドブックの情報拡張を行う．

1. はじめに

旅行者が，旅先の観光情報を収集するために利用する情報源の一つとして，旅行ガイドブックが挙げられる．「るるぶ」のような旅行ガイドブックは，一般的に観光地ごとに発行され，有名な観光名所，土産物，宿泊施設，飲食店など，観光に関連する基本的な情報が掲載されている．観光情報を収集するための他の情報源としては，旅行会社や地方公共団体が運営する観光ポータルサイトが挙げられるが，観光地により情報量に大きな差があり，長い期間更新されないままのサイトもある．そのため，旅先の基本的な観光情報を得るために，まずは旅行ガイドブックを手にとってみる，というユーザも少なくない．しかし，具体的に旅行の計画を行う際には，旅行ガイドブックに多数掲載されている飲食店のどのお店を利用すればよいのか，家族連れでも快適に過ごすには，どの宿泊施設を選択すればよいか判断に迷う場面が多々ある．このような場合には，過去に同じ観光地を旅行した旅行者の経験は大いに役に立つ情報である．過去の旅行者の経験を収集するための情報源として，ブログ著者が旅行での体験を記述した旅行ブログエントリ，旅行に関連する知識や知恵を教え合う場である質問応答コンテンツが挙げられる．

そこで，本研究では，観光地に関する基本的な情報がまとめて掲載されている旅行ガイドブックに対し，旅行ブログエントリや質問応答コンテンツを自動的に対応付けることで旅行ガイドブックの情報拡張を行う手法を提案する．また，自動的に対応付けられた結果を閲覧できるシステムの構築を行う．本システムでは，ユーザが iPad などのタブレット上で電子化された旅行ガイドブックを閲覧する際に，旅行ブログエントリや質問応答コンテンツを同時に閲覧できるシステムである．このシステムを利用することで，基本的な観光情報は旅行ガイドブックから，また，旅行者の豊かな経験に基づく多様な情報は，対応付けられた旅行ブログエントリや，質問応答コンテンツから得ることができる．本研究で作成したシステムは，旅行ガイドブックへ情報を付与するため，拡張現実の一例とみなすことができる．

本論文の構成は以下の通りである．2 節ではシステムの動作例，3 節では関連研究，4 節では提案手法，5 節では実験結果と考察について述べ，6 節で本稿をまとめる．

2. システム動作例

本研究で構築したシステムについて，その動作例を紹介する．図 1 は，本研究の提案手法で情報拡張された旅行ガイドブックの例でり，屋久島・奄美・種子島に関する旅行ガイドブックの中で，加計呂麻島に関するページである[*1]．このページには，加計呂麻島の見所や，宿泊施設に関する情報が記載されている．「ブログ」ボタン（図中①）をクリックすると，旅行ガイドブックに関連する旅行ブログエントリが閲覧できる．また，「知恵袋」ボタン（図中②）をクリックすると，旅行ガイドブックに関連する質問応答コンテンツを閲覧することができる．図 2 は，図 1 の旅行ガイドブックに対応付けられた質問応答コンテンツの一例であり，加計呂麻島の宿泊施設に関する質問と，その回答が記述されている．質問者は，おすすめの民宿について質問しており，民宿に泊まるであれば加計呂麻島よりもうけ島が，また，家族で楽しむのであれば渡連や諸数のペンションが良い，と回答者が薦めている．この例からもわかるように，本研究で作成したシステムでは，基本的な観光情報は旅行ガイドブックから，また，旅行者の豊かな経験に基づく多様な情報は，旅行ガイドブックに対応付けられた旅行ブログエントリや，質問応答コンテンツから得ることができる．

1 広島市立大学大学院　情報科学研究科
Graduate School of Information Sciences, Hiroshima City University
2 広島市立大学　情報科学部
School of Information Sciences, Hiroshima City University

a) ishino, nanba, kobayashi, takezawa}@ls.info.hiroshima-cu.ac.jp

*1 るるぶ「屋久島 奄美 種子島 '09～10」, JTB パブリッシング, pp.70-71 (2009).

図 1　情報拡張された旅行ガイドブックの例

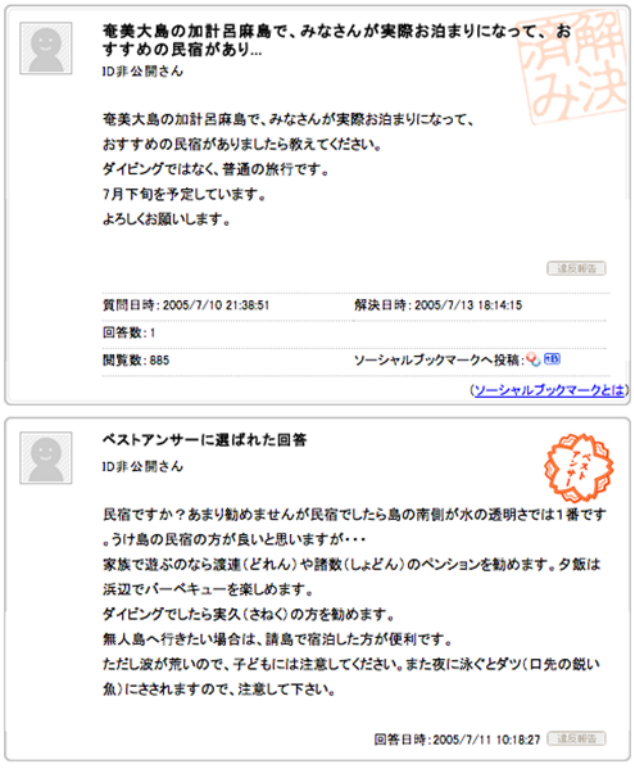

図 2　図 1 の旅行ガイドブックに自動的に対応付けられた質問応答コンテンツの例

3. 関連研究

本節では，本研究に関連する研究について述べる．まず，書籍に Web 上の情報を自動的に付与する研究について説明する．Rakesh ら[1]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 Web サイトから検索，収集し，教科書に付与する手法を提案している．Rakesh らの研究と本研究では，書籍に Web 上の情報を付与する点では同じであるが，本研究は，書籍として旅行ガイドブック，付与する情報として旅行ブログエントリ，質問応答コンテンツを使用する点で異なる．

本研究は，旅行ガイドブックに，旅行ブログエントリや質問応答コンテンツを付与する手法を提案しているが，本研究は，ある文書に対し，関連文書を付与する研究の一例とみなすことができる．ある文書に対し，関連文書を自動的に付与する研究の例として，コンテンツ連動型広告に関する研究がある[2，3]．コンテンツ連動型広告とは，Web ページの文脈や重要語を抽出し，内容の関連性の高い広告を配信するシステムである．

本研究では，旅行ガイドブックの情報拡張のために，旅行ブログエントリと質問応答コンテンツを利用する．旅行ブログやそのエントリを登録したポータルサイトとしては，“Travel Blog”[*2]，“旅行・観光ブログ村”[*3]，“フォートラベル”[*4]などがある．これらのポータルサイトでは，ブロガーが自身のブログを旅行ブログとして登録することで，

*2 http://www.travelblog.org/
*3 http://www.travelblog.org/
*4 http://www.travelblog.org/

旅行ブログの集積を行う．しかし，ブログ空間にはたくさんのブログが存在するため，このようなポータルサイトに登録されていない一般ブログの中にも，旅行ブログエントリが多数存在する．そこで，Nanba ら[4]は，一般ブログから，機械学習を使用して旅行ブログエントリを自動的に検出する手法を提案している．機械学習の手法には，CRF を採用している． Nanba らは，上記の手法により，精度 86.7%と高い精度で旅行ブログエントリの検出に成功している．本研究では，Nanba らの手法により収集した旅行ブログエントリを使用する．

また，近年，ブログ著者の属性(性別，年齢，居住域など)を文体や記載内容から自動的に推定する研究が進んでいる[5, 6, 7]．このような技術を利用することで，本研究で構築したシステムのユーザと，似た属性を持つブログ著者が記述した旅行ブログエントリを優先的に提示することで，ユーザに適した観光情報の推薦ができるようになると考えられる．

質問応答コンテンツの代表例としては，“Yahoo!知恵袋”[*5]，“OKWave”[*6]などがある．本研究では，旅行ガイドブックに対応付ける質問応答コンテンツとして，「地域，旅行，お出かけ」カテゴリに登録されている Yahoo!知恵袋を利用する．対応付の際には，質問応答コンテンツのタイプ分類を行う．本研究と同様に，質問応答コンテンツのタイプを分類する研究がある．渡邊ら[8]の研究では質問応答コンテンツにおいて，回答者の質問の選択を容易にすることを目的に，以下の 5 つの質問タイプに分類する方法を提案している．分類には SVM を使用している．渡邊らは，5 つのタイプに分類を行うが，本研究では観光に特化した分類を行うため，4.2 節で定義するタイプに分類を行う．

- 事実（事象の定義，真実，客観的な理由や手段を問う質問）
- 根拠（客観的な根拠，理由を問う質問）
- 経験（回答者の経験や体験がなければ回答できない質問）
- 提案（情報提供を依頼や問題の解決方法を問う質問
- 意見（推測，嗜好など主観的に回答をしてよい質問）

本研究と同様に，観光の支援を目的とした研究がある．旅行の計画を立てる際に，旅行先でのイベントに関する情報や，観光名所をどのような順序で訪れるかといった行動経路は大変重要な情報である．Web からイベント情報を抽出する研究[9，10，11]や，行動経路を抽出する研究がある[12，13，14]．本研究では，旅行ガイドブックへ，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が，上記の研究により抽出された，イベント情報や，行動経路の情報を旅行ガイドブックに対応付けることで，旅行ガイドブックの情報を更に拡張ができる可能性がある．

*5 http://chiebukuro.yahoo.co.jp/
*6 http://okwave.jp/

4. 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

本節では，旅行ブログエントリと質問応答コンテンツを旅行ガイドブックに対応付けることで旅行ガイドブックの情報拡張を行う手法について説明を行う．本研究では，各コンテンツ内に含まれる単語を利用して，対応付けを行う手法を提案する．最終的な目標は，旅行ガイドブックの各ページに，旅行ブログエントリや，質問応答コンテンツを対応付けることである．広島に関する旅行ガイドブックを見ると，各ページには，広島に関連する情報が記載されているが，「広島」という単語が必ず含まれるわけではない．この場合，ページ単位に，広島と各ページに関連する旅行ブログエントリや質問応答コンテンツを対応付ける事は困難である．そのため，本研究では，まず，旅行ブログエントリと質問応答コンテンツを旅行ガイドブック単位で対応付ける．この処理を行うことで，広島に関連する旅行ブログエントリや，質問応答コンテンツを収集できると考えられる．旅行ガイドブック単位での対応付け手法については，4.1 節で説明を行う．

次に，旅行ブログエントリと質問応答コンテンツを旅行ガイドブックのページ単位で対応付ける．本研究では，旅行ガイドブックの中でレストランについて言及しているページには，その観光地における飲食関連の情報について記述した旅行ブログエントリや質問応答コンテンツを対応付けることを目的とする．そのためには，各コンテンツが，どのようなタイプの情報について言及しているのかを分類する必要がある．このタイプ分類の手法については，4.2 節で説明を行う．

4.1 旅行ガイドブック単位での対応付け

本節では，旅行ブログエントリと質問応答コンテンツを旅行ガイドブック単位で対応付ける手法について説明を行う．旅行ガイドブックには，紹介されている観光地の名前，旅行ブログエントリには，ブログ著者が訪れた観光地の名前，質問応答コンテンツには，質問者の質問のターゲットとなっている観光名所の名前が頻繁に出現する．そのため，対応付けを行う際に，各コンテンツに出現する「地名」は大きな手掛かりになると考えられる．よって，本研究では，各コンテンツに含まれる地名の出現頻度を使用することで，旅行ガイドブックと旅行ブログエントリ，旅行ガイドブックと質問応答コンテンツの類似度を計算し，対応付けを行う手法を提案する．地名の抽出には，日本語構文解析器 CaboCha[*7]を使用する．旅行ガイドブックでは，旅行ガイドブック一冊に含まれる地名の頻度を利用する．旅行ブログエントリと質問応答コンテンツの地名の出現頻度の求め方は，以下の手法を採用する．

*7 http://code.google.com/p/cabocha/

旅行ブログエントリの地名の出現頻度の求め方

- LOC：タイトルと本文に含まれる地名の頻度.
- LOC+TITLE : タイトルと本文に含まれる地名の頻度. ただしタイトルに含まれる地名については頻度を 10 とする.

質問応答コンテンツの地名の出現頻度の求め方

- LOC(Q)：質問応答コンテンツの質問文に含まれる地名の頻度.
- LOC(Q&A) : 質問応答コンテンツの質問文と回答文に含まれる地名の頻度.

上記の手法により，地名の出現頻度を求め，類似度の計算を行う. 類似度の計算には, 汎用連想計算エンジン GETA [*8]を用いる．本研究では，以下の手法により，旅行ブログエントリと質問応答コンテンツを，旅行ガイドブックへ対応付ける.

- F-base : 対応付けを行う際に利用する類似度の閾値は，訓練用データで F 値が最も高くなる値を採用し，類似度が閾値以上のガイドブックへ対応付けを行う.
- TOP1：類似度が最も高いガイドブックへ対応付けを行う.
- 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へ対応付けを行う.

TOP1 と THRE は, 精度を重視した対応付け手法である. 精度を重視した対応付け手法を採用したのは，旅行ブログエントリや質問応答コンテンツは，日々大量に作成されているため，精度重視の対応付けを行っても，十分な数の旅行ブログエントリや質問応答コンテンツを，旅行ガイドブックへ対応付けることがでるためである.

4.2 タイプ分類

本節では，各コンテンツをタイプ分類する手法について説明を行う. 旅行ガイドブックは, 観光名所に関する情報, お土産物に関する情報，宿泊施設に関する情報などタイプごとに分けて情報が掲載される傾向がある．本研究では，旅行ガイドブックを分析し，各コンテンツを 6 種類のタイプに分類することとした．分類するタイプと，その判定基準を表 1 に示す．1 ページ内に，「見る」と「買う」に関する情報が記載されている場合は，タイプは,「見る」と「買う」両方に判定される．このような判定を行うことで，複数のタイプの情報が記載されている旅行ガイドブックに対しても，適切なタイプの旅行ブログエントリや質問応答コンテンツを対応付けることができる.

*8 http://geta.ex.nii.ac.jp/geta.html

表 1　タイプと判定基準

タイプ	判定基準
見る	観光名所などの見るだけで楽しめる物についての情報を記載されている.
体験する	○○体験やスキューバダイビングなど，自分の体を使って楽しめる物についての情報が記載されている.
買う	お土産に関する情報が記載されている.
食べる	飲食に関する情報が記載されている.
泊まる	宿泊に関する情報が記載されている.
その他	「見る」,「体験する」,「買う」,「食べる」,「泊まる」に該当しない場合．例として広告ページや巻末の交通情報.

本研究では，機械学習により各コンテンツのタイプの判定を行う．学習には,「手掛かり語の有無」を素性として与える．「体験する」に分類されたページには,「スキューバダイビング」や「シュノーケリング」といったアクティビティの種類,「買う」に分類されたページには,「お土産」や「スイーツ」といった単語がよく使われる．このような語を人手で収集し，手掛り語として使用する．使用する手掛り語を表 2 に示す.

表 2　手掛かり語の例

タイプ	手掛かり語の例	件数
見る	世界遺産，博物館，夜景，ミュージカル，オペラ，入館　など	48
体験する	スキューバダイビング，シュノーケリング，パラセイリング　など	144
買う	お土産，ブランド，スイーツ，免税店，買う　など	165
食べる	食べる，レストラン，ランチ，ディナー，グルメ　など	36
泊まる	泊まる，ロッジ，客室，チェックイン，チェックアウト　など	22

本研究では，機械学習に手掛かり語を素性して与えることで，タイプ分類を行う手法を提案している．しかし，旅行ガイドブックには，観光地の画像が大きく紹介されており，文字情報が少ないページがあるため，機械学習に与える素性が手掛かり語のみでは，正しくタイプ分類が行えない場合がある．そのため，本研究では，旅行ガイドブックのタイプ分類には，手掛かり語に加え，色情報を素性として使用する．旅行ガイドブックの「見る」に分類されたページを分析すると，例えば，美しい海や，景色が紹介されているページが多くあった．このようなページには，青空や海の画像が大きく掲載されており，全体的に青い色が多

い印象を受ける.そのため,文字情報の少ないページでも，旅行ガイドブックのページの色情報を利用することで，タイプ分類が行えるようになると考えられる.

本研究では，色情報として，カラーヒストグラムを採用する．カラーヒストグラムとは，カラー画像における RGB 値の分布を示したものである．本研究では，旅行ガイドブック 1 ページを 1 枚の画像とみなし，格子状の各点から色情報を抽出し，カラーヒストグラムを求め，それらを機械学習の素性に用いる．一般的に，各点の RGB は，0-255 までの 256 階調で表現されるが，このままでは，機械学習に与える素性が膨大になってしまう．このため，RGB は，256 階調から 4 階調に減らしてカラーヒストグラムを作成し，色情報として 64 個(4 の 3 乗)の素性を機械学習に与える.

5. 実験

5.1 旅行ガイドブック単位での対応付け

実験に用いるデータ

OCR によりテキスト化した観光ガイドブック 90 冊に対し，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．旅行ブログエントリとしては，Nanba らの手法により収集した旅行ブログエントリ 489 件，質問応答コンテンツとしては,「地域，旅行，お出かけ」カテゴリに登録されている Yahoo!知恵袋 1990 件に対し，人手により判定を行った結果を実験に使用した．評価尺度には，精度，再現率，F 値を使用した.

比較手法

本研究では，各コンテンツに含まれる地名を利用して，類似度を計算する手法を提案した．この提案手法の有効性を確認するため，各コンテンツに含まれる内容語（名詞，動詞，形容詞）の頻度を類似度の計算に利用する以下の3種類の手法を比較手法として，実験を行った.

- BASELINE1：旅行ブログエントリに含まれる内容語.
- BASELINE2：質問応答コンテンツの質問文に含まれる内容語の頻度.
- BASELINE3：質問応答コンテンツの質問文と回答文に含まれる内容語の頻度.

実験結果と考察

F-base 手法で，旅行ガイドブックと旅行ブログエントリの対応付けを行った結果を表 3 に，旅行ガイドブックと質問応答コンテンツの対応付けを行った結果を表 4 に示す．表 3，4 より，比較手法に比べ，いずれの提案手法でも，精度，再現率，F 値が向上した．よって，本研究で提案した，各コンテンツに含まれる地名により類似度を算出し，旅行ガイドブックに対応付ける手法の有効性が確認できたといえる.

表 3　旅行ブログエントリの対応付けの結果

	精度(%)	再現率(%)	F 値(%)
BASELINE1	12.7	16.5	14.1
LOC	54.4	44.2	48.7
LOC+TITLE	56.6	42.8	48.4

表 4　質問応答コンテンツの対応付けの結果

	精度(%)	再現率(%)	F 値(%)
BASELINE2	24.9	24.7	24.6
BASELINE3	24.7	24.3	24.4
LOC(Q)	65.3	28.9	38.6
LOC(Q+A)	50.5	38.4	42.3

表 3，表 4 の実験結果で，最も精度の高い LOC+TITLE, LOC(Q)の手法により地名の出現頻度を求め，TOP1 手法, THRE 手法を用いて対応付けた結果を，表 5, 表 6 に示す．表 5，表 6 より，TOP1 手法よりも，THRE 手法の方が高い精度を得ることができた．また，THRE 手法において，旅行ブログエントリの対応付けでは，77.4%，質問応答コンテンツの対応付けでは，92.4%と高い精度が得られた．THRE 手法を用いた場合，旅行ガイドブック 1 冊に対して旅行ブログエントリ 99 件，質問応答コンテンツ 1561 件が対応付けられており，再現率の低さは問題ないといえる.

表 5　精度重視の旅行ブログエントリの対応付けの結果

	精度(%)	再現率(%)	F 値(%)
TOP1	71.7	30.6	42.9
THRE	77.4	22.7	35.1

表 6　精度重視の質問応答コンテンツの対応付けの結果

	精度(%)	再現率(%)	F 値(%)
TOP1	85.2	16.7	28.0
THRE	92.4	8.1	14.9

精度重視で行った THRE 手法により対応付けを失敗した原因としては，主に，旅行ガイドブックに含まれる交通手段に関するページに記載されている地名が原因であった．旅行ガイドブックの交通手段のページには，様々な大都市や空港，駅から観光地へ行く方法が記載されているため，旅行ガイドブックがターゲットとして紹介している観光地以外の地名が多数出現する．そのため，本研究の提案手法では，正しく対応付ができなかった．この問題を解決するためには，4.2 節で提案した旅行ガイドブックのタイプ分類を先に行い，交通情報が記載されたページが分類される「その他」に分類されたページに含まれる地名を，類似度の計算から除くことで，正しく対応付けを行うことができるようになると考えられる.

5.2 タイプ分類

実験に用いるデータ

OCR によりテキスト化した観光ガイドブック 20 冊分である 2897 ページを実験対象とした．旅行ブログエントリとしては，Nanba らの手法により収集した旅行ブログエントリ 1001 件，質問応答コンテンツとしては，「地域，旅行，お出かけ」カテゴリに登録されている Yahoo!知恵袋 1500 件に対し，人手によりタイプ分類を行った結果を実験に使用した．人手によりタイプ分類を行った結果を表 7 に示す．

表 7　人手によりタイプ分類した結果

タイプ	旅行ガイドブック	旅行ブログエントリ	質問応答コンテンツ
見る	1026	465	633
体験する	78	293	391
買う	418	188	191
食べる	741	450	530
泊まる	273	161	245
その他	365	56	0

機械学習と評価尺度

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．評価尺度として精度，再現率を使用した．

実験手法

本研究では，機械学習に与える素性として，旅行ガイドブックのタイプ分類においては，手掛り語と色情報，旅行ブログエントリと質問応答コンテンツのタイプ分類においては手掛り語を使用する手法を提案した．素性として手掛り語のみを用いた手法を提案手法 1，手掛り語と色情報を用いた手法を提案手法 2，色情報のみを用いた手法を提案手法 3 として実験を行った．また，提案手法の有効性を確認するため，各コンテンツに含まれる全単語を素性として使用した場合を，比較実験として行った．

実験結果と考察

旅行ガイドブックのタイプ分類の結果を表 8 に，旅行ブログエントリのタイプ分類の結果を表 9 に，質問応答コンテンツのタイプ分類の結果を表 10 に示す．

旅行ガイドブックのタイプ分類の結果について考察する．表 8 より，比較手法に比べ，提案手法 1 では，高い精度を得ることができた．しかし，旅行ガイドブックには，画像が大きく掲載され，文字情報が少ないページがあるため，

表 8　旅行ガイドブックのタイプ分類の結果

素性	評価尺度	見る	体験する	買う	食べる	泊まる	平均
全単語	精度(%)	46.0	16.9	25.1	41.7	39.1	33.7
	再現率(%)	52.6	15.4	23.3	49.2	17.0	31.5
提案手法 1 (手掛かり語)	精度(%)	69.0	84.2	70.3	77.4	69.7	74.1
	再現率(%)	34.6	25.1	9.3	36.8	26.2	26.4
提案手法 2 (手掛かり語＋色)	精度(%)	67.6	37.2	34.6	66.1	52.9	51.7
	再現率(%)	59.3	35.0	24.1	49.9	42.5	42.2
提案手法 3 (色)	精度(%)	61.0	-	-	56.4	-	-
	再現率(%)	37.2	-	-	4.6	-	-

表 9　旅行ブログエントリのタイプ分類の結果

素性	評価尺度	見る	体験する	買う	食べる	泊まる	平均
全単語	精度(%)	69.2	55.8	42.2	72.9	41.7	56.4
	再現率(%)	67.7	40.6	28.8	63.2	33.0	46.7
提案手法 1 (手掛かり語)	精度(%)	72.6	100.0	58.3	76.7	81.7	77.8
	再現率(%)	58.9	55.0	20.7	62.5	29.9	45.4

表 10　質問応答コンテンツのタイプ分類の結果

素性	評価尺度	見る	体験する	買う	食べる	泊まる	平均
全単語	精度(%)	71.1	52.8	63.2	67.4	71.4	65.2
	再現率(%)	69.8	46.1	38.6	64.5	60.4	55.9
提案手法 1 (手掛かり語)	精度(%)	76.4	66.5	75.1	75.4	78.4	74.4
	再現率(%)	62.8	14.5	44.2	51.1	61.1	46.8

手掛り語のみを素性として使用した提案手法 1 では，タイプ分類を行うことができないページが多くあった．そのため，提案手法 1 の再現率は低くなってしまった．この問題を解決するために，提案手法 2 と提案手法 3 では，色情報を素性として利用した実験を行った．

色情報のみを素性として使用した提案手法 3 では，「見る」の場合においてのみ，手掛かり語のみを使用した提案手法 2 と同程度の精度と再現率を得ることができた．人手で「見る」に分類された旅行ガイドブックを分析すると，屋外で撮影された画像が多く用いられていた．屋外で撮影された画像には，空や海が多く含まれており，青い色が多く出現していたため，色情報がタイプ分類に有効に働いたと考えられる．そのため，素性に手掛かり語と色情報を使用した提案手法 2 では，「見る」において，高い精度を保ったまま，再現率を向上させることができた．

提案手法 2 の「見る」以外では，提案手法 1 に比べ，再現率は改善できたが，精度は大きく低下した．よって，今後は，「見る」に関しては，提案手法 2 を，「見る」以外のタイプに関しては，提案手法 1 を採用することが考えられる．また，提案手法 2 では，色情報のみを画像から抽出した情報として利用したが，今後は，Bag of Visual Words[15]など, 画像の分類によく使用される手法を利用することで，精度，再現率ともに改善できると考えられる．

旅行ガイドブックと質問応答コンテンツのタイプ分類の結果について考察する. 表 9, 表 10 より, 比較手法に比べ，提案手法 1 で，高い精度を得ることができたが，再現率は低下した．再現率低下の主な原因は，手掛かり語の不足であった．本研究では，タイプ分類に人手で収集した手掛り語を使用したため，手掛かり語の数が十分でなく，タイプ分類ができず再現率が低下した．この問題は，文書分類などの分野で，トピックに特有な語を抽出する際に利用される，トピックモデル[16]を利用し，手がかり語を増やすことで解決できると考えられる．

6. おわりに

本研究では，旅行ブログエントリと質問応答コンテンツを，旅行ガイドブックに対応付けるため，旅行ブログエントリと質問応答コンテンツを旅行ガイドブック単位で対応付ける手法と，タイプ分類を行う手法を提案した．また，提案手法の有効性を確認するため，実験を行った．旅行ガイドブックと旅行ブログエントリの対応付では 77.4%，旅行ガイドブックと質問応答コンテンツの対応付では 92.4%と，高い精度での対応付けに成功した．また，タイプ分類において, 旅行ガイドブックでは精度 74.1%, 再現率 42.2%, 旅行ブログエントリでは精度 77.8%，再現率 45.4%，質問応答コンテンツでは精度 74.4%，再現率 46.8%を得た．また，実験により得られた結果を使用し，情報拡張された旅行ガイドブックを閲覧できるシステムを構築した．

謝辞

JTB パブリッシングの発行する旅行ガイドブックるるぶを使わせて頂いたことに深く御礼申し上げます．

参考文献

[1] Rakesh, A., Sreenivas, G., Anitha, K. and Kishnaram, K.: Enriching Textbooks with Images, Proc. 20th ACM Conference on Information and Knowledge Management (2011).
[2] Broder, A., Fontoura, M. and Josifovski, V.: A Semantic Approach to Contextual Advertising, Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.559-566 (2007).
[3] Ishino, A., Nanba, H., and Takezawa, T.: Providing Ad Links to Travel Blog Entries Based on Link Types, Proc. 9th Workshop on Asian Language Resources collocated with IJCNLP 2011, pp. 63-70 (2011).
[4] Nanba, H., Taguma, H., Ozaki, T., Kobayashi, D., Ishino, A. and Takezawa, T.: Automatic Compilation of Travel Information from Automatically Identified Travel Blogs, Proc. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing, Short Paper, pp.205-208 (2009).
[5] Yasuda, N., Hirao, T., Suzuki, J. and Isozaki, H.: Identifying Bloggers' Residential Areas, Proc. AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp.231-236 (2006).
[6] Ikeda, D., Takamura, H. and Okumura, M.: Semi-supervised Learning for Blog Classification, Proc. 23rd AAAI Conference on Artificial Intelligence, pp.1156-1161 (2008).
[7] Schler, J., Koppel, M., Argamon, S. and Pennebaker, J.: Effects of Age and Gender on Blogging, Proc. AAAI Symposium on Computational Approaches for Analyzing Weblogs, pp.199-205 (2006).
[8] 渡邊直人, 島田諭, 関洋平, 神門典子, 佐藤哲司: QA コミュニティにおける質問者の期待に基づく質問分類に関する一検討, 電子情報通信学会, 第 3 回データ工学とマネジメントに関するフォーラム, B5-1 (2011).
[9] 岡本昌之, 菊池匡晃: ブログからの地域イベント情報抽出, 情報処理, Vol.51, No. 1, pp.14-17 (2010).
[10] 藤坂達也, 李龍, 角谷和俊: 地域イベント発見および特性検証のための実空間マイクロブログを用いたユーザ移動パターン分析システム, 情報処理学会創立 50 周年記念(第 72 回)全国大会, pp.845-846 (2010).
[11] 斉藤隆太, 石野亜耶, 難波英嗣, 竹澤寿幸: 新聞記事と Web からのイベント情報の自動抽出, 電子情報通信学会第 20 回 Web インテリジェンスとインタラクション研究会 (2011).
[12] 郡宏志, 服部峻, 手塚太郎, 田島敬史, 田中克己: ブログからのビジターの代表的な経路とそのコンテキスト抽出, 情報処理学会研究報告データベースシステム研究会, Vol.2006, No.78, pp.35-42 (2006).
[13] Davidov, D.: Geo-mining: Discovery of Road and Transport Networks Using Directional Patterns, Proc. 2009 Conference on Empirical Methods in Natural Language Processing, pp.267-175 (2009).
[14] Ishino, A., Nanba, H., and Takezawa, T.: Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries, Proc. 18th international Conference on Information Technology and Travel &Tourism (2011).
[15] Yang, J., Jiang Y.G., Hauptmann, A. and Ngo, C.W.: Evaluating Bag-of-Visual-Word Representation in Scene Classification, Proc. International Workshop on Workshop on Multimedia Information Retrieval, pp.197-206 (2007).
[16] 立川華代, 小林一郎: 文書から取得した制約知識に基づく潜在的トピック抽出, 言語処理学会第 18 回年次大会, pp.313-316 (2012).