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† 藤井 一輝† 藤原 泰士† 前田 剛† 難波 英嗣† 竹澤 寿幸†

†広島市立大学大学院 情報科学研究科 〒731-3194 広島市安佐南区大塚東 3 丁目 4 番 1 号

E-mail: †{ishino, fujii, fujiwara, maeda, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

あらまし 観光を支援する媒体のひとつとして，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，様々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，旅行での体験を記述した旅行ブログエントリや，旅行に関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，これらのコンテンツを旅行ガイドブックへ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目指す．有効性を確認するための実験を行い，旅行ブログエントリでは 0.820，質問応答コンテンツでは 0.770 の割合で，適切に対応付けを行うことができた．また，実験で得られた結果を使用して，情報拡張された旅行ガイドブックを閲覧するシステムを構築した．

キーワード 観光情報処理, 旅行ガイドブック, ブログ, 質問応答コンテンツ

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．株式会社 JTB パブリッシングが出版している「るるぶ」などの旅行ガイドブックには，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，具体的に旅行を計画する際には，旅行ガイドブックに多数掲載されている飲食店の中で，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，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のページに対し，関連する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手法を提案し，旅行ガイドブックの情報を拡張する．また，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の構築を行う．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，過去の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．そのため提案システムは，旅行の計画を行う際に，有用なシステムであると言える．

本論文の構成は以下の通りである．2 節ではシステムの概要および動作例，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. システムの概要および動作例

本節では，本研究で構築したシステムの概要，および動作例について説明する．まず，システムの概要を述べる．本研究で構築したシステムでは，紙媒体の旅行ガイドブックをスキャンし，OCR（光学式文字読取装置）処理したものを入力すると，旅行ガイドブックの各ページに対し，関連する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．

次に，提案する対応付け手法を実装し，構築したシステムの動作例を紹介する．本システムは，iPad などのタブレット端末での閲覧を想定している．図 1 は，提案手法により情報拡張された旅行ガイドブックのページの例である．図 1 は，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，旅行ガイドブックのページに対応付けられた旅行ブログエントリを閲覧できる．また，「知恵袋」ボタン（図中②）をクリックすると，旅行ガイドブックのページに対応付けられた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のページ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，民宿に泊まるのであれば加計呂麻島よりうけ島が，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，本研究で構築したシステムでは，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．提案システムでは，旅行ガイドブックのページに対し，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ているため，上記の例のように，旅行ガイドブックのページに掲載されている

[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).

① ②

図 1: 情報拡張された旅行ガイドブックのページの例

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

複数の観光名所や宿泊施設の比較や感想が記述されてい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ことが可能である．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は，複数の観光名所や宿泊施設に関連する情報の他に，以下の様な情報を得ることができると考えられる．

- 旅行ガイドブックに含まれないローカルな情報
- 季節や，天候に応じたお勧めの観光情報
- 一人旅などの旅行形態に応じた観光情報

3. 関連研究

本節では，本研究に関連する研究，サービスを紹介する．本研究では，書籍である旅行ガイドブックのページに，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手法を提案している．本研究と同様に，書籍に，Web 上の情報を自動的に付与する研究がある．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 Wikipedia から検索，収集し，対応付ける手法を提案している．Rakesh らは，教科書に画像を対応付けることで，視覚的な情報を補うことを目的としている．本研究では，旅行ガイドブックのページに対応付ける情報として，旅行ブログエントリと質問応答コンテンツを採用し，過去の旅行者の経験を付与することを目的としている点で異なる．NDL ラボ[2]では，脚注表示機能を有した電子読書支援システムの構築実験[3]を行っている．電子読書支援システムの構築実験では，OCR により，書籍からテキスト情報を抽出し，そのテキストに含まれる Wikipedia 日本語版のタイトルを検出し，Wikipedia 内の写真と説明文を，書籍の左右のサイドノートに表示するシステムの開発を行っている．電子読書支援システムの構築実験では，ページ内のキーワードに対し，関連する Wikipedia のページを参照しているが，本研究では，旅行ガイドブックのページに対し，旅行ガイドブックと質問応答コンテンツを対応付けており，より対象とするページに関連のある情報を対応付けることができると考えられる．

本研究では，旅行ガイドブックのページに対応付ける情報として，旅行ブログエントリと質問応答コンテンツを使用している．Nanba ら[2]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，精度 0.867 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

2 http://lab.kn.ndl.go.jp/cms/
3 http://lab.kn.ndl.go.jp/nii/

法により収集した旅行ブログエントリを使用する．質問応答コンテンツの代表例としては，Yahoo! 知恵袋[4]，OKWave[5]などがある．本研究では，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!知恵袋を使用する．

本研究では，旅行ガイドブックのページ，旅行ブログエントリ，質問応答コンテンツのタイプ分類を行い，その結果を，旅行ガイドブックのページへの，旅行ブログエントリと質問応答コンテンツの対応付けに利用する．本研究では，藤井ら[3]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のタイプ分類を行う．藤井らの研究のように，文書のタイプ分類には，文書中のテキスト情報が利用されている．本研究では，旅行ガイドブックのページに対してもタイプ分類を行うが，旅行ガイドブックには，テキスト情報の他に，旅行先の景色，お土産，ホテルなどの画像が多数掲載されているという特徴がある．神谷ら[4]は，欧 10 都市の旅行ガイドブックを対象に，掲載されている画像の構成要素について分析を行い，単体の建造物や広場，橋などが多く掲載されていることを明らかにしている．そのため，旅行ガイドブックに掲載されている画像の構成要素（画像情報）は，タイプ分類において重要な素性の一つになると考えられる．本研究では，画像情報として，Bag of Visual Words [5]を使用する．Bag of Visual Words とは，1 つの画像から複数の局所特徴をベクトル量子化してヒストグラム化したものであり，近年，物体認識技術において最もよく使用される技術である[6]. Yang ら[7]は，Bag of Visual Words により画像情報を抽出し，画像の分類を行う手法を提案している．本研究では，テキスト情報に，画像情報を加えることで，旅行ガイドブックのページのタイプ分類を行うことを目的としているため，Yang らの研究と異なる．

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

本研究では，旅行ガイドブックのページへ，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手法を提案する．提案手法の流れを以下に示す．Step 1，Step 2，Step 3 について，それぞれ 4.1 節，4.2 節，4.3 節で説明を行う．

- Step 1 旅行ブログエントリのページ，旅行ブログエントリ，質問応答コンテンツのタイプ分類．
- Step 2 旅行ガイドブックへの，旅行ブログエントリと質問応答コンテンツの対応付け．
- Step 3 旅行ガイドブックのページへの，旅行ブログエントリと質問応答コンテンツの対応付け．

4.1. 旅行ガイドブックのページ・旅行ブログエントリ・質問応答コンテンツのタイプ分類

本研究では，旅行ガイドブックのページに，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手法を提案する．まず，対応付けを行う旅行ガイドブックのページの分析を行う．一般的に，旅行ガイドブックではページごとに，観光名所に関する情報，土産物に関する情報，宿泊施設に関する情報など，表 1 に示す観光に特化したタイプにまとめられて掲載されている．

[4] http://chiebukuro.yahoo.co.jp
[5] http://okwave.jp/

表 1: 旅行ガイドブックのページのタイプとその内容

タイプ	内容
見る	観光名所などの見て楽しめる物やイベントについての情報．
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．
買う	土産物に関する情報．
食べる	飲食に関する情報．
泊まる	宿泊施設に関する情報．
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．

そのため，旅行ガイドブックのページへ，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際には，旅行ガイドブックのページ，旅行ブログエントリ，質問応答コンテンツのタイプを判定し，旅行ガイドブックのページと同じ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で，自然な対応付けができると考えられる．図 3 に示すように，「宮島」のガイドブックのタイプ「見る」に判定されたページには，同じタイプ「見る」の旅行ブログを対応付けると，自然な対応付けができる．

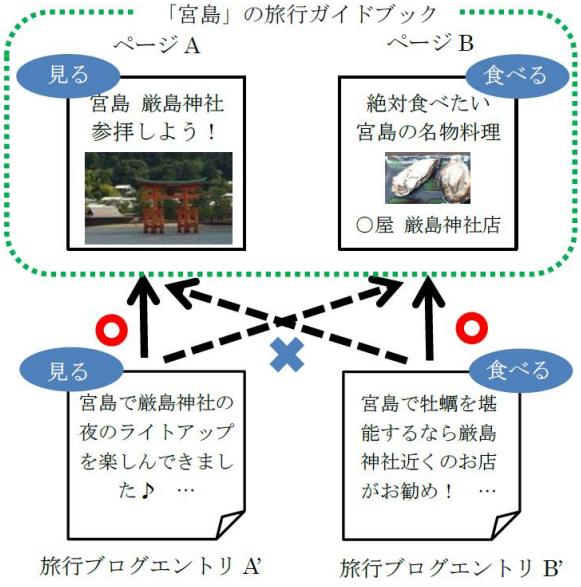

図 3: 旅行ガイドブックのページへの旅行ブログエントリの対応付けのイメージ

そこで本研究では，旅行ガイドブックの 1 ページ，旅行ブログエントリ，質問応答コンテンツを，表 1 に示すタイプのうち，「その他」を除く「見る」，「体験する」，「買う」，「食べる」，「泊まる」の 5 種類のタイプに分類し，対応付けに利用する．旅行ガイドブックの 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」両方に分類する．旅行ブログエントリ，質問応答コンテンツも同様に分類する．このような分類を行うことで，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のページに対しても，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

ると考えられる．本研究では，旅行ガイドブックのページのタイプ分類には，テキスト情報と画像情報を用いる．旅行ブログエントリ，質問応答コンテンツのタイプ分類には，テキスト情報を使用する．

■　テキスト情報を使用したタイプ分類

藤井ら[3]は，旅行ブログエントリのテキスト情報を使用することで，本研究と同じタイプに分類する手法を提案している．本研究では，テキスト情報を使用したタイプ分類には，藤井らの手法を利用する．藤井らは，タイプごとに情報利得を利用することで，各タイプに特有の単語を手掛かり語として収集し，手掛かり語の有無を機械学習に使用している．藤井らの手法により旅行ガイドブックから収集した手掛かり語の例を表 2 に示す．旅行ブログエントリ，質問応答コンテンツの各タイプにおいても，手掛かり語を収集する．

表 2: 旅行ガイドブックから収集した手掛かり語の例

タイプ	手掛かり語の例
見る	展示，見る，見学，みどころ
体験する	インストラクター，初心者，体験
買う	アイテム，揃う，小物，ブランド
食べる	食べる，味わえる，料理，シェフ
泊まる	宿，ロビー，部屋，内風呂男女
その他	記入，航空券，航空会社，申告

■　画像情報を使用したタイプ分類

旅行ガイドブックには，多数の画像が掲載されている．タイプ「見る」に判定された旅行ガイドブックのページには，海や山など景色の画像が多く掲載されている．また，タイプ「食べる」に判定されたページには，料理の画像が多く掲載されている．そのため，旅行ガイドブックのページに，どのような画像が含まれているかという情報は，タイプ分類において重要な手掛かりになると考えられる．よって，旅行ガイドブックのページのタイプ分類には，手掛かり語の有無に加え,画像情報を機械学習の素性に用いる.本研究では，画像情報として，Bag of Visual Words[5]を使用する．

本研究では，まず，訓練用の画像集合から，Dense sampling により局所特徴を抽出し，局所特徴をクラスタリングすることで代表ベクトル（Visual word）を作成する．クラスタリングには K-means を利用し，1000 個の Visual word を作成する.タイプ分類を行う旅行ガイドブックのページに対して，近似する Visual Word の出現回数をカウントし，ヒストグラムを作成することで Bag of Visual Words を作成する．本研究では，旅行ガイドブックのページごとに Bag of Visual Words を作成し，機械学習の素性として与える．

4.2. 旅行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

本研究では，旅行ガイドブックのページへ，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ことを目的としている．しかし，広島に関する旅行ガイドブックを分析すると，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，旅行ガイドブックのページへ，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と考えられる．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，旅行ガイドブックに対応付ける．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.類似した処理を行う研究に，He ら[8]らの研究がある．He らは，論文中の文脈を一部与えるとその文脈に即した関連論文を検索し，自動的に推薦する研究を行っている．関連研究を検索する際のキーワードに，論文全体に関連するキーワードとしてタイトルやアブストラクトから抽出したキーワード（global context）と，関連研究を付与する文脈に出現するキーワード（local context）を使用することで，検索精度が向上することを報告している.本研究では，「広島」などの旅行ガイドブック全体に関連するキーワードが global context，対応付けるページに出現するキーワードが local context に相当する．本研究においても，対応付けの精度向上のため，まず，旅行ブログエントリと質問応答コンテンツを，旅行ガイドブックに対応付け（global context による対応付け），次に，その旅行ガイドブックのどのページに対応付けるか判定を行う（local context による対応付け）．

旅行ガイドブックでは，紹介されている観光地の名前，旅行ブログエントリでは，ブログ著者が訪れた観光地の名前，質問応答コンテンツでは，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，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に出現する「地名」は重要な手掛かりになると考えられる．よって，本研究では，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，旅行ガイドブックへ，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．各コンテンツからの地名の抽出には，日本語構文解析器 CaboCha[6]を使用する．

また，旅行ガイドブックへの対応付けの際に，Step 1 で判定した，旅行ガイドブックのページ，旅行ブログエントリ，質問応答コンテンツのタイプ分類の結果を使用する．タイプ分類の結果を，旅行ガイドブックへの対応付けに使用する意義を述べる．テキストは，その種類に応じて，特徴的な構成要素を持つことが知られている．例えば，学術論文では,「背景」,「目的」,「方法」,「結論」,「考察」などの特徴的な構成要素がある．Kando[9]の研究では，論文検索のタスクにおいて，論文の構成要素を解析し，特定の意味役割のみの文を使用して index を作成した方が，論文全文を使うよりも検索精度が高くなることを報告している．旅行ガイドブックにおける構成要素は，表 1 に示すタイプである．そこで本研究では，構成要素としてタイプ分類の結果を利用することで，高精度の対応付けを目指す．タイプ分類の結果を使用した旅行ガイドブックへの対応付けの流れを,図 4 に示す.図 4 に示すように，タイプ「体験する」に分類された旅行ブログエントリを対応付ける旅行ガイドブックを選択する際には，その旅行ブログエントリから抽出された地名と，旅行ガイドブックのタイプ「体験する」に判定されたページのみから抽出された地名を使用する．他のタイプの場合においても同様の操作を行う．

[6] http://code.google.com/p/cabocha/

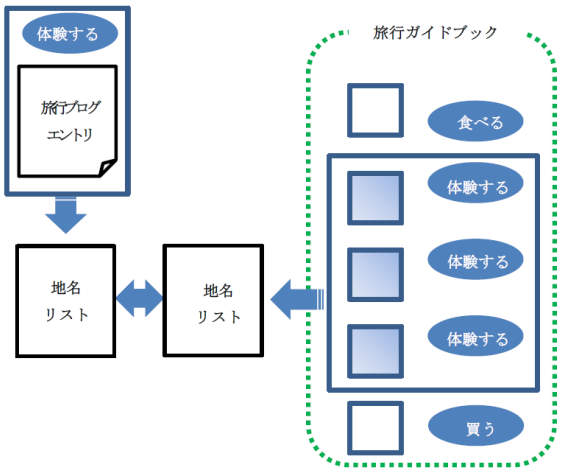

図 4: タイプを使用した旅行ガイドブックへの旅行ブログエントリの対応付け

本研究では，抽出した地名リストを使用して，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を対応付ける．対応付けの手法として，k 近傍法を使用した手法（KNN 手法）と，機械学習を使用した手法（SVM 手法）を提案する．

■ k 近傍法を使用した手法（KNN 手法）

KNN 手法では，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求め，閾値より高い類似度を持つ場合に，対応付けを行う．類似度の計算には SMART[10]を使用する．閾値は，予備実験の結果から求める．

■ 機械学習を使用した手法（SVM 手法）

機械学習には，SVM を使用する．旅行ガイドブックごとに学習器を構築し，対象の旅行ブログエントリや質問応答コンテンツが，対応付けられるか，対応付けられないのかという 2 値分類を行う．

4.3. 旅行ガイドブックのページ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

本節では，旅行ブログエントリと質問応答コンテンツを，旅行ガイドブックのページに対応付ける手法を説明する．Step 2 より，旅行ブログエントリと質問応答コンテンツが対応付けられる旅行ガイドブックは判定済である．対象となる旅行ブログエントリや質問応答コンテンツと，同じタイプを持つ旅行ガイドブックのページとの類似度を，地名の出現頻度を使用して求め，最も類似度の高い旅行ガイドブックのページに対応付ける．類似度の計算には，コサイン類似度を使用する．単語の重みには，TF*IDF を使用する．IDF は，Web 検索エンジンにおけるヒット件数を用いる．

5. 評価実験

本研究で提案した手法の有効性を確認するため，以下の 3 種類の実験を行った．実験の詳細については，それぞれ，5.1 節，5.2 節，5.3 節で述べる．本研究で構築したシステムの有用性の評価を 5.4 節で行った．

- 旅行ガイドブックのページ，旅行ブログエントリ，質問応答コンテンツのタイプ分類
- 旅行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
- 旅行ガイドブックのページ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

本実験では，旅行ブログとして，Nanba ら[2]の手法により収集した旅行ブログエントリ，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!知恵袋を使用する．

5.1. 旅行ブログエントリのページ，旅行ブログエントリ，質問応答コンテンツのタイプ分類

5.1.1. 実験条件

【実験に用いるデータ】 OCR 処理を行った旅行ガイドブック 2897 ページ（20 冊分），旅行ブログエントリ 1000 件，質問応答コンテンツ 1500 件を使用した．上記のデータに対し，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 3 に示す．

表 3: 人手によるタイプ判定の結果

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	102	395	620
体験する	78	241	412
買う	418	163	191
食べる	741	382	502
泊まる	278	134	257

【機械学習と評価尺度】 機械学習を用いてタイプ分類を行った．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】 旅行ブログエントリ，質問応答コンテンツのタイプ分類には，情報利得を利用して収集した手掛かり語を素性として与える藤井ら[3]の手法を使用した（IG 手法）．旅行ガイドブックのタイプ分類では IG 手法に加え，以下に示す画像情報（Bag of Visual Words）を使用する手法について実験を行う．

- IG+BoVW: 情報利得を利用して収集した手掛かり語と，画像情報を素性として与える．
- BoVW：画像情報を素性として与える．

5.1.2. 実験結果と考察

藤井らの手法により，旅行ブログエントリのタイプ分類では平均で精度 0.636，再現率 0.467，質問応答コンテンツのタイプ分類では平均で精度 0.810，再現率 0.315 を得た．旅行ガイドブックのページのタイプ分類の結果を，それぞれ表 4 に示す．

旅行ガイドブックのページのタイプ分類での IG 手法，IG+BoVW 手法，BoVW 手法の実験結果について考察を行う．IG 手法と IG+BoVW 手法は，平均では同程度の結果であった．しかし，タイプ「見る」においては，精度は同程度のまま，再現率を 5.2 ポイント向上させることに成功した．BoVW 手法においても，タイプ「見る」では，精度 0.619，再現率 0.147 を得ることができた．IG+BoVW 手法は，IG 手法と比較し，マクネマー検定により有意水準 0.050 で統計的に有意であることがわかった．よって，タイプ「見る」の判定において，画像情報は有効であると言える．これは，タイプ「見る」に判定されたページには，海や山などの景色の写真が多用されており，有効な画像情報を取

表 4: 旅行ガイドブックのページのタイプ分類の結果

手法	評価尺度	見る	体験する	買う	食べる	泊まる	平均
IG	精度	0.733	0.917	0.815	0.805	0.740	0.802
	再現率	0.321	0.170	0.200	0.324	0.328	0.269
IG+BoVW	精度	0.741	0.917	0.762	0.776	0.754	0.800
	再現率	0.373	0.170	0.287	0.353	0.368	0.302
BoVW	精度	0.619	—	—	0.720	—	—
	再現率	0.147	—	—	0.056	—	—

り出しやすかったためではないかと考えられる．タイプ分類におけるテキスト情報と画像情報の有効性の確認のため，表 4 のタイプ「見る」について，旅行ガイドブックのページに含まれる文字数ごとに精度，再現率を算出した結果を表 5 にまとめる．表 5 における文字数 100 は，旅行ガイドブックのページに含まれる文字数が 0~100 文字であるページを使用した場合であり，文字数 200 は，旅行ガイドブックのページに含まれる文字数が 101~200 文字であるページを使用した場合を示している．

表 5: タイプ「見る」における文字数ごとの実験結果

文字数	IG 手法		IG+BoVW 手法	
	精度	再現率	精度	再現率
100	0.000	0.000	0.679	0.034
200	0.627	0.181	0.667	0.200
300	0.250	0.094	0.286	0.125
400	0.563	0.294	0.585	0.311
500	0.417	0.240	0.417	0.240

表 5 より，タイプ「見る」において，200 文字以下の文字数が少ない場合においては，IG 手法に比べ，IG+BoVW 手法では，精度・再現率ともに高い結果を得ることができており，画像情報が有効な素性として働いていることがわかった．しかし，文字数が増えると，画像情報を加えることでの有意差は小さくなっている．現在は，旅行ガイドブックの 1 ページを画像として扱うことで，Bag of Visual Words を作成しているため，Bag of Visual Words を構築する際に，文字も画像を構成する要素として取りこまれる．そのため，旅行ガイドブックの各ページに文字が多く含まれていると，画像情報を取り出す際のノイズになっていると考えられる．山口ら[11]は，文書画像から，文字領域や図表などの領域を自動的に分離するための手法を提案している．山口らの研究成果を利用することで，旅行ガイドブックから，文字が記載されている領域と画像領域を分割することができれば，IG+BoVW 手法の実験結果を改善することができると考えられる．

5.2. 旅行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

5.2.1. 実験条件

【実験に用いるデータ】 旅行ガイドブック 90 冊，旅行ブログエントリ 918 件，質問応答コンテンツ 1998 件を使用した．被験者には，旅行ガイドブックと質問応答コンテンツを閲覧し，類似度の高い旅行ガイドブックに対応付けるよう指示した．評価尺度として精度，再現率を使用した．

【実験手法】 提案手法の有効性を確かめるため，以下に示す 3 種類の提案手法と，2 種類の比較手法について実験を行った．KNN_TYPE 手法においてのみタイプ分類の結果を使用し，他の手法ではタイプ分類の結果は考慮せず，旅行ガイドブックの全てのページから抽出した地名を実験に使用した．

＜提案手法＞

- KNN_TYPE : KNN 手法を用いる．類似度の計算には，地名の出現頻度を使用する．タイプ分類の結果を考慮し，旅行ガイドブックからは，旅行ブログエントリや質問応答コンテンツと同じタイプに判定されたページのみから地名を抽出する．
- KNN_LOC : KNN 手法を用いる．類似度の計算には，地名の出現頻度を使用する．
- SVM_LOC : SVM 手法を用いる．素性に地名の出現頻度を使用する．

＜比較手法＞

- BASE_KNN : KNN 手法を用いる．類似度の計算には，名詞の出現頻度を使用する．
- BASE_SVM : SVM 手法を用いる．素性に名詞の出現頻度を使用する．

5.2.2. 実験結果と考察

旅行ガイドブックと旅行ブログエントリの対応付けの実験結果を表 6，旅行ガイドブックと質問応答コンテンツの対応付けの結果を表 7 に示す．表 6，表 7 より，SVM 手法より，KNN 手法の方が，高い精度を得ることができた．KNN 手法では，名詞の出現頻度を用いる BASE_KNN 手法よりも，地名の出現頻度を用いる KNN_TYPE 手法，KNN_LOC 手法の方が高い精度を得ることができた．また，タイプ分類の結果を使用しない KNN_LOC 手法に比べ，タイプ分類の結果を使用する KNN_TYPE 手法では，旅行ブログエントリでは 4.4 ポイント，質問応答コンテンツでは 7.5 ポイント精度を改善することができており，最も高い精度を得ることができた．KNN_TYPE 手法は，精度は高いが，再現率は低かった．しかし，KNN_TYPE 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．よって，本研究の提案手法の有効性を確認できたといえる．

表 6: 旅行ブログエントリの対応付け結果

実験手法		精度	再現率
比較手法	BASE_KNN	0.273	0.155
	BASE_SVM	0.216	0.030
提案手法	KNN_TYPE	0.811	0.204
	KNN_LOC	0.767	0.201
	SVM_LOC	0.440	0.190

表 7: 質問応答コンテンツの対応付け結果

実験手法		精度	再現率
比較手法	BASE_KNN	0.487	0.166
	BASE_SVM	0.405	0.184
提案手法	KNN_TYPE	0.858	0.210
	KNN_LOC	0.783	0.206
	SVM_LOC	0.398	0.301

対応付けの失敗の主な原因としては，旅行ブログエントリや，質問応答コンテンツから，旅行の目的地以外の地名が抽出されてしまうことが挙げられる．旅行ブログエントリでは，ブログ著者が旅行として訪れた場所の情報の他に，自宅から旅行先への経路を詳細に記述する場合がある．その場合には，旅行先の地名だけでなく，自宅近くの地名や，移動の間に訪れた場所の地名が抽出される．また，質問応答コンテンツでは，「京都から，東京まで遊びに行こうと思いますが，新幹線代がちょっと高くて気になります！新幹線よりもう少し安く東京まで行く方法を教えてください.」の様に，移動元の情報が記述されていると，旅行先ではない地名が出現する．旅行ブログエントリから，旅行者の行動経路を抽出する研究として，Ishino ら[12]の研究がある．Ishino らの研究では，旅行ブログエントリから，機械学習を用いて，「地名」→「地名」に移動した，などのような，旅行者の行動経路を自動で抽出する手法を提案している．Ishino らの手法を，旅行ブログエントリや，質問応答コンテンツに適用することで，旅行の目的地を抽出し，目的地以外の地名を削除できると考えられる．

5.3. 旅行ガイドブックのページ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

5.3.1. 実験条件

【実験に用いるデータ】 旅行ブログエントリ 100 件，質問応答コンテンツ 100 件に対し，旅行ガイドブック 90 冊分のページへの対応付け実験を行った．

【実験手法】 以下に示す 2 種類の提案手法と比較手法について実験を行った．提案手法では，5.2 節の KNN_TYPE 手法により対応付けられる旅行ガイドブックが判定済みである．比較手法では，旅行ガイドブックへの対応付けを行わず，旅行ブログエントリ，質問応答コンテンツと，旅行ガイドブックのページとのコサイン類似度を求め，最も類似度の高い旅行ガイドブックへのページへ対応付ける．

- 提案手法 1：対応付けられた旅行ガイドブック内でコサイン類似度を求め，もっとも類似度の高いページに対応付ける．ページへの対応付けの際に，タイプ判定の結果は使用しない．そのため，旅行ガイドブックのページと異なるタイプの旅行ブログエントリや質問応答コンテンツが対応付けられる場合がある．
- 提案手法 2：対応付けられた旅行ガイドブック内でコサイン類似度を求め，もっとも類似度の高いページに対応付ける．ページへの対応付けの際には，タイプ判定の結果を使用する．そのため，旅行ガイドブックのページと同じタイプの旅行ブログエントリや質問応答コンテンツが対応付けられる．

【評価方法】 本研究で提案した手法の有効性を確認するため，提案手法 1，提案手法 2，比較手法の 3 つの手法により得られた対応付け結果に対し，アンケート調査を行った．アンケート調査の被験者は，大学生と大学院生の 11 名である．旅行ガイドブックのページに対応付けられた旅行ブログエントリと質問応答コンテンツに対し，それぞれ被験者から，対応付けが「適切である」または，「適切でない」の 2 件法で回答を得て，過半数以上の被験者が「適切である」と回答した対応付けを，「適切である」と判定した．

5.3.2. 実験結果と考察

表 8 に，旅行ブログエントリ，質問応答コンテンツに対し，対応付け結果が「適切である」と判定された割合を示す．旅行ブログエントリの実験結果においては，比較手法に比べ，提案手法 1，2 がよい結果を得ることができた．よって，比較手法のように，旅行ガイドブックのページと旅行ブログエントリの対応付けを一度に行うのではなく，提案手法 1，2 のように，まずは対応付ける旅行ガイドブックを決定し，その旅行ガイドブック内のページに対応付けを行う方が，適切に対応付けを行うことができることを示せたといえる．提案手法 1 と提案手法 2 に差が見られなかったのは，旅行ブログエントリは，記述量が多いものが多く，複数のタイプに分類される旅行ブログエントリも多いため，対応付けの際に差が生じなかったと考えられる．

質問応答コンテンツでは，提案手法 2 が最もよい結果を得た．質問応答コンテンツでは，「○○でのお勧めのレストランはありますか？」などのように，食事や宿泊施設などタイプを絞った質問が行われるため，旅行ガイドブックのページへの対応付けを行う際に，タイプ分類の結果を利用する提案手法 2 が有効に働いたと考えられる．

表 8: 対応付けが「適切である」と回答された割合

実験手法	旅行ブログエントリ	質問応答コンテンツ
比較手法	0.720	0.530
提案手法 1	0.820	0.570
提案手法 2	0.820	0.770

5.4. システムの有用性評価

本研究では，提案した手法により情報拡張された旅行ガイドブックを閲覧するシステムを構築した．構築したシステムが，旅行の計画を行う際に有用であるかどうかについて，被験者 11 名に対し，アンケート調査を行った．その結果を，図 5 に示す．なお，「1: まったくそうは思わない」，「2: そうは思わない」と回答した被験者は 0 名であった．図 5 より，旅行ガイドブックや質問応答コンテンツを使用し，情報拡張された旅行ガイドブックは，旅行計画の際に有用であるといえる．

被験者による自由記述からは，提案システムを使用することで，旅行ガイドブックに掲載されていないような，旅行の経験を活かした情報や，季節や天候，旅行形態に応じた情報を得ることができたという回答を得ることができた．しかし，一方で，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が長文であり，欲しい情報を得るのに時間がかかるといった問題点が

ある．これは，旅行ブログエントリの要約を作成することで改善できると考えられる．また，ユーザの好みや，旅行の形態に合わなければ，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は役に立たないといった回答もあった．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[13, 14, 15]．このような研究の成果を利用することで，本研究で構築したシステムの利用者と，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，ユーザに適した旅行ブログエントリや質問応答コンテンツの推薦ができるようになると考えられる．また，旅行ブログエントリの著者や，質問応答コンテンツの質問者の旅行の際の条件（季節，天候，旅行形態など）をそれぞれのテキスト情報から抽出することで，よりシステム利用者の状況に即した旅行ブログエントリや質問応答コンテンツの推薦が可能になると考えられる．利用者の条件に即した，旅行ガイドブックの情報拡張は，今後の研究の課題である．

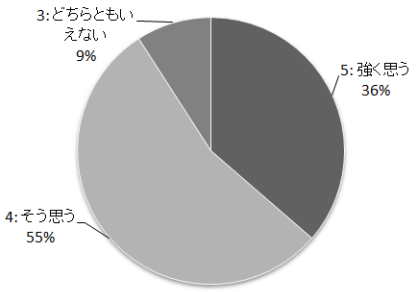

図 5: 情報拡張された旅行ガイドブックを閲覧するシステムの有用性評価

6. おわりに

本研究では，旅行ガイドブックへ，旅行ブログエントリ，質問応答コンテンツを対応付けることで，旅行ガイドブックの情報を拡張する手法を提案した．提案手法は 3 ステップからなる．Step 1 では，旅行ブログエントリのページ，旅行ブログエントリ，質問応答コンテンツのタイプ分類を行った． Step 2 では，旅行ガイドブックと旅行ブログエントリ，質問応答コンテンツの対応付けを行った．Step 3 では，旅行ガイドブックのページと旅行ブログエントリ，質問応答コンテンツの対応付けを行った．有効性を確認するための実験を行い，旅行ブログエントリでは 0.820，質問応答コンテンツでは 0.770 の割合で適切に対応付けを行うことができた．また，構築したシステムに対し評価実験を行い，旅行の計画を行う際に，提案システムが有効であることを示した．

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, pp.1847-1856 (2011).

[2] 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).

[3] 藤井一輝，石野亜耶，藤原泰士，前田剛，難波英嗣，竹澤寿幸: 多言語旅行ブログエントリを用いた観光情報提示システム，第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2014), 2014.

[4] 神谷文子，浦山益郎，北原理雄: 主題要素の写され方からみた都市景観写真の構図に関する研究 欧 10 都市の観光ガイドブックを事例として，日本建築学会計計画論文集，Vol.528，pp.179-186 (2000).

[5] Csurka, G., Dance, C. R., Fan, L., Willamowski, J. and Bray, C.: Visual Categorization with Bags of Keypoints, Proc. ECCV International Workshop on Statistical Learning in Computer Vision, pp.1-22 (2004).

[6] 柳井啓司: 一般物体認識の現状と今後，情報処理学会論文誌, Vol.48, No.SIG 16 (CVIM 19), pp.1-24 (2007).

[7] 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).

[8] He, Q., Pei, J., Kifer, D., Mitra, P. and Giles, C.L.: Context-aware Citation Recommendation, Proc. World Wide Web Conference 2010, pp.421-430 (2010).

[9] Kando, N.: Text-level Structure of Research Papers: Implications for Text-Based Information Processing Systems, Proc. British Computer Society Annual Colloquium of Information Retrieval Research, pp.68-81 (1997).

[10] Salton, G: The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ (1971).

[11] 山口拓真，丸山稔: 確率的トピックモデルによる文書画像の領域分割(画像認識,コンピュータビジョン)，電子情報通信学会論文誌．D，Vol.J92-D, No.6, pp.876-887 (2009).

[12] 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, pp.113-124 (2011).

[13] 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).

[14] Ikeda, D., Takamura, H. and Okumura, M.: Semi-supervised Learning for Blog Classification, Proc. 23rd AAAI Conference on Artificial Intelligence, pp.1156-1161 (2008).

[15] 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).