原著論文

Original Paper

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

Enriching Travel Guidebooks with Travel Blog Entries and Archives of Answered Question

石野 亜耶　広島市立大学大学院　情報科学研究科
Aya Ishino　Graduate School of Information Sciences, Hiroshima City University.
ishino@ls.info.hiroshima-cu.ac.jp, http://aya-info.main.jp/

藤井 一輝　（同上）
Kazuki Fujii　fujii@ls.info.hiroshima-cu.ac.jp

藤原 泰士　（同上）
Taishi Fujiwara　fujiwara@ls.info.hiroshima-cu.ac.jp

前田 剛　（同上）
Tsuyoshi Maeda　maeda@ls.info.hiroshima-cu.ac.jp

難波 英嗣　（同上）
Hidetsugu Nanba　nanba@hiroshima-cu.ac.jp, http://www.ls.info.hiroshima-cu.ac.jp/~nanba/

竹澤 寿幸　（同上）
Toshiyuki Takezawa　takezawa@hiroshima-cu.ac.jp

Keywords: travel information processing, travel guidebook, blog, archives of answered question

Summary

Travelers planning to visit a particular tourist spot need information about their destination and they often use travel guidebooks (guidebooks) to collect this information. However, guidebooks lack specific information, such as first-hand accounts by users who have visited the specific destination. To compensate for the lack of such information, we focused on travel blog entries (blog entries) and archives of answered question (QA archives). In this paper, we propose a method for enriching guidebooks by aligning with blog entries and question answering archives. This is a three-step method. In Step1, we classify pages of guidebooks, blog entries and QA archives into five types of content, such as “watch” and “eat.” In Step 2, we align each blog entry and QA archive with guidebooks by taking these content types into account. In Step 3, we align each blog entry and QA archive with individual pages in guidebooks. To investigate the effectiveness of our method, we conducted a few experiments. Accordingly, 82.0% of blog entries and 77.0% of QA archives were judged to be helpful for travelers. Finally, we constructed a prototype system that provides enriched guidebooks.

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

図 1　情報拡張された旅行ガイドブックのページの例

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

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

*[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).

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．

旅行者の過去の経験を収集する場としては，楽天トラベル*[2]のような宿泊施設の予約サイトや，食べログ*[3]のような飲食店の口コミサイトがある．このようなサイトでは，一件の宿泊施設や飲食店に対し，その個別の施設に関する口コミしか得ることができない．そのため，本研究では，旅行ガイドブックに対応付ける情報源として，旅行ブログエントリと質問応答コンテンツを採用した．

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は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

本研究の関連研究として，文書の情報拡張に関する研究，旅行ブログエントリや質問応答コンテンツに関連するサービスや研究，タイプ分類，観光支援に関する研究を紹介する．

3.1 文書の情報拡張

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[Rakesh 11]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*[4]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*[5]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[Broder 07，Ishino 11a]．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で，旅行者の情報収集の支援を行うことを目的としているため異なる．

3.2 旅行ブログエントリや質問応答コンテンツに関連するサービスや研究

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た．旅行ブログエントリや質問応答コンテンツに関連するサービスや研究を紹介する．

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*[6]，旅行・観光ブログ村*[7]，

*2 http://travel.rakuten.co.jp/
*3 http://tabelog.com/
*4 http://lab.kn.ndl.go.jp/cms/
*5 http://lab.kn.ndl.go.jp/nii/
*6 http://www.travelblog.org/
*7 http://travel.blogmura.com/

フォートラベル*[8]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，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[Nanba 09]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 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．

質問応答コンテンツの代表例としては，Yahoo! 知恵袋*[9]，OKWave*[10]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!知恵袋を使用する．

3.3 タイプ分類

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．本研究と同様に，旅行ブログエントリや質問応答コンテンツを自動分類する研究がある．徳久ら[徳久 11]は，ブログエントリから，観光開発のためのヒントを抽出するために，ブログエントリ中の文に対し，ヒント文であるか，ヒント文でないのかを，自動で分類する手法を提案している．渡邊ら[渡邊 11]は，回答者の質問の選択を容易にすることを目的に，質問応答コンテンツの質問を，「事実」（事象の定義，真実，客観的な理由や手段を問う質問）や「根拠」（客観的な根拠，理由を問う質問）など，5 種類の質問タイプに自動分類する手法を提案している．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に対し，4.1 節で定義する観光に特化したタイプに分類する点で異なる．また，徳久らや渡邊らは，タイプ分類を研究の主な目的としているが，本研究では，タイプ分類の結果を，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点で異なる．

上記の研究のように，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，旅行先の景色や，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[神谷 00]は，欧米 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 [Csurka 04]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ている技術である[柳井 07]. Yang ら[Yang 07]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．Yang らは，Bag of Visual Words を利用し，画像自体の分類を目的としているのに対し，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．

3.4 観光支援

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[岡本 10, 藤坂 10, 斉藤 12]や，行動経路を抽出する研究がある[群 06, Davidov 09, Ishino 11b]．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できる可能性がある．

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う．本研究では，旅行ガイドブックのページと同じタイプの情報が記載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，旅行ガイドブックのページに対応付ける．タイプ分類の結果は，Step 2 以降で行う対応付けに利用する．

Step 2 旅行ガイドブックのブック単位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．本研究では，旅行ガイドブックのページに，旅行ブロ

*8 http://4travel.jp/
*9 http://chiebukuro.yahoo.co.jp
*10 http://okwave.jp/

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．その前段階として，本ステップでは，旅行ブログエントリと質問応答コンテンツを，どの旅行ガイドブックに対応付けるのかを判定する．

Step 3 旅行ガイドブックのページ単位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．Step 2 において，対応付ける旅行ガイドブックは判定済のため，本ステップでは，その旅行ガイドブックのどのページに対応付けるかを判定する．Step 3 では，Step 1 と Step 2 での実験結果を利用する．そのため，Step 3 での旅行ガイドブックのページ単位への旅行ブログエントリと質問応答コンテンツを対応付けた結果が，提案システムの出力結果となる．

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の対応付けを行う手法を提案する．まずは，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．

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	土産物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	宿泊施設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．

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 は，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である．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」の旅行ブログエントリを対応付けると，自然な対応付けができる．

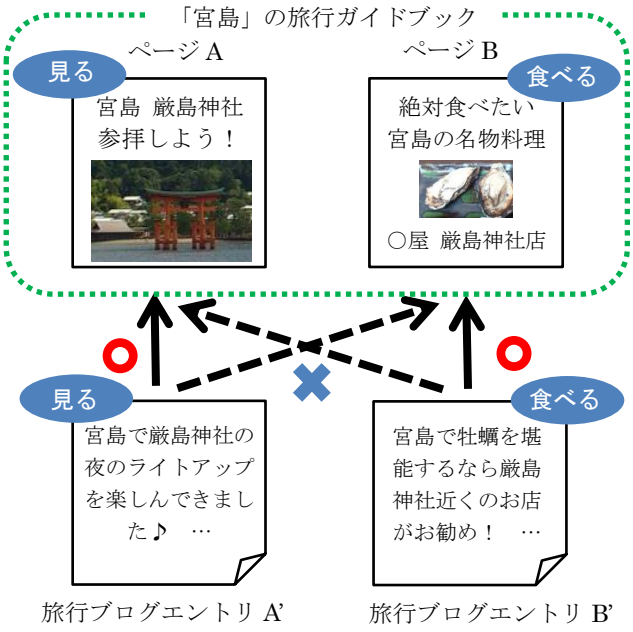

図 3 旅行ガイドブックのページへの旅行ブログエントリの対応付けのイメージ

旅行ガイドブックのページに掲載されている情報は，主にタイプ「見る」,「体験する」,「買う」,「食べる」,「泊まる」である．本研究では,「見る」,「体験する」,「買う」,「食べる」,「泊まる」に該当しないページを「その他」と定義している．タイプ「その他」には，広告ページや観光地に到着するまでの交通情報，海外の旅行ガイドブックでは，パスポートの取得方法など，旅行ガイドブックが対象としている観光地に直接関係しないページが多く含まれる．そのため，本研究では，タイプ「見る」,「体験する」,「買う」,「食べる」,「泊まる」のいずれかのタイプに判定される旅行ガイドブックのページに，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と質問応答コンテンツを対応付ける手法を提案する．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ページには，あらかじめタイプごとに情報が記載されているが，旅行ガイドブックは毎年更新されるため，人手でのタイプ分類は現実的ではない．そのため，旅行ガイドブックのページに対しても，自動でタイプ分類を行う．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，旅行ブログエントリ，質問応答コンテンツのタイプを，機械学習を用いて分類する．機械学習には，TinySVM を用いる．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素性に用いる．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

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，「展示」，「見学」などの単語が頻繁に出現する．また，タイプ「体験する」に判定された旅行ガイドブックのページには，「インストラクター」，「体験」などの単語が頻繁に出現する．このように，各タイプに判定されたページには，そのタイプに特有の単語が頻出する傾向がある．また，タイプ「見る」，「体験する」，「買う」，「食べる」，「泊まる」に判定されない旅行ガイドブックのページに特有な単語として，タイプ「その他」に特有な単語を使用することができる．そのため，本研究では，「見る」，「体験する」，「買う」，「食べる」，「泊まる」，「その他」の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の素性として与える．本研究では，手掛かり語の収集は，情報利得により自動で収集する．情報利得を利用することで，手掛かり語を収集するためのコストを抑えると同時に，より素性として有効な手掛かり語を収集することができると考えられる．

本研究では，「見る」，「体験する」，「買う」，「食べる」，「泊まる」，「その他」の 6 種類のタイプごとに，出現する単語に対して，情報利得を求める．情報利得を求める際に使用する単語は，MeCab*[11]により分割された形態素とし，品詞が名詞句，動詞，形容詞であるものとする．またこれらの単語のうち，出現回数が 1 回以下，単語の長さが 15 文字以上，単語の長さが半角 1 文字以下のいずれかに当てはまる単語は，不用語として削除する．上記の条件にあてはまる単語に対し，情報利得を求め，その値が，閾値より高い単語を手掛かり語として収集する．閾値は，予備実験により設定した．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，同様に，手掛かり語を収集する．

表 2 旅行ガイドブックから情報利得により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，博物館
体験する	インストラクター，初心者，体験，自然
買う	アイテム，揃う，店内，小物，ブランド
食べる	食べる，店，味わえる，料理，シェフ
泊まる	宿，ロビー，部屋，空間，内風呂男女
その他	記入，航空券，航空会社，原則，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 を使用する．Bag of Visual Words は，画像を局所特徴の出現頻度ベクトルで表したものであり，一般物体認識のタスクにおいて，広く普及している画像特徴表現である．Bag of Visual Words は，自然言語処理の分野で，単語の出現頻度ベクトルで文書を表現する Bag of Words を，画像へ応用したものである．

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．

旅行ブログエントリにも，多数の画像が含まれている場合があるため，旅行ブログエントリのタイプ分類においても，画像情報は重要な手掛かりのひとつになると考えられる．本研究では，Yahoo!ブログ*[12]から収集した旅行ブログエントリを実験対象としている．Yahoo!ブログから旅行ブログエントリを収集した時点では，ブログ中の画像の所在は Javascript で埋め込まれており，クローラを使用しての画像の自動収集が困難であったため，本研究では，旅行ブログエントリのタイプ分類には，テキスト情報のみを用いる．

*[11] http://mecab.sourceforge.net/

*[12] http://blogs.yahoo.co.jp/

4.2 旅行ガイドブックのブック単位への対応付け

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[He 10]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）．

本節では，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を対応付ける手法について説明する．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*[13]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[Kando 97]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．

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）を提案する．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[Salton 71]を使用する．2 分割交差検定により，訓練用のデータで，再現率が 15.0%以上を保ち，最も精度が高くなる値を閾値とする．

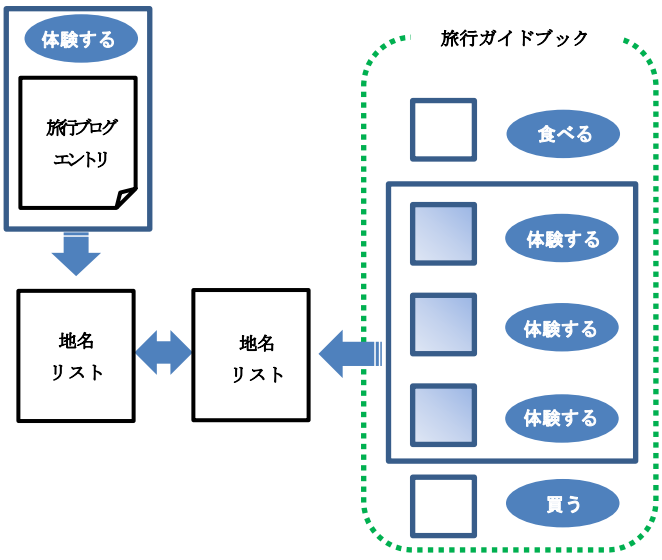

図 4 タイプを使用した旅行ガイドブックへの旅行ブログエントリの対応付け

*[13] http://code.google.com/p/cabocha/

4.3 旅行ガイドブックのページ単位への対応付け

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ことで求める．

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．また，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．

(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
(2) 旅行ガイドブックのブック単位への対応付け
(3) 旅行ガイドブックのページ単位への対応付け

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

§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)，Nanba らの手法により収集した旅行ブログエントリ 1000 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 9388 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．旅行ガイドブックは，1 ページごとにタイプ分類を行った．また，人手による判定においても，1 ページに複数のタイプの情報が記載されている場合には, 複数のタイプに判定した．旅行ブログエントリ，質問応答コンテンツにおいても，同様に判定した．人手によりタイプ分類を行った結果を表 3 に示す．

表 3 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の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	1026	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
その他	365	56	7888

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．タイプ分類の結果は，Step2 以降の実験に使用するため，タイプ分類の精度が低いと，全体のシステムの性能の低下に繋がる．また，旅行ブログエントリや質問応答コンテンツは日々作成されるため，Web 上には大量のデータが存在している．そのため，再現率は低くとも，使用する旅行ブログエントリと質問応答コンテンツの件数を増やすことで，旅行ガイドブックに対応付ける旅行ブログエントリと質問応答コンテンツの件数は増やすことが可能であると考えられる．上記の理由から，再現率よりも精度を重視する．

実験手法

以下に示す提案手法について実験を行った．また，提案手法の有効性を確認するため，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に含まれる全単語を素性として使用した場合を，比較手法として実験した．

＜提案手法＞

- IG：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．
- 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（Bag of Visual Words）を素性として与える．
- BoVW：画像情報（Bag of Visual Words）を素性として与える．

＜比較手法＞

- Word：全単語を素性として与える．

§実験結果と考察

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，それぞれ表 4，表 5，表 6 に示す． 表 4 の BoVW 手法で，機械学習により学習できなかった部分は，「—」と記載した．学習できなかった原因は，旅行ガイドブックのデータ不足であると考えられる．

まず，提案手法である IG 手法と，比較手法である Word 手法の実験結果について，考察を行う．実験結果より，旅行ガイドブックのページでは 28.7 ポイント，質問応答コンテンツでは 10.1 ポイント精度を向上させることができた. 旅行ブログエントリでは, タイプ「見る」以外のタイプでは精度の向上に成功している．詳細は後述するが，旅行ガイドブックのページのタイプ「見る」の分類では，画像情報を加えた IG+BoVW 手法で精度を向上させることができることを確認している．旅行ブログエントリには，画像情報が含まれることが多いため，今後，IG+BoVW 手法により精度の改善が可能であると考えられる．よって，情報利得により収集した手掛かり

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
Word（比較手法）	精度(%)	46.0	16.9	25.1	41.7	39.1	46.9
	再現率(%)	53.6	15.4	23.3	49.2	17.0	30.9
IG（提案手法）	精度(%)	73.3	91.7	81.5	80.5	74.0	75.6
	再現率(%)	32.1	17.0	20.0	32.4	32.8	27.9
IG+BoVW（提案手法）	精度(%)	74.1	91.7	76.2	77.6	75.4	75.8
	再現率(%)	37.3	17.0	28.7	35.3	36.8	33.7
BoVW（提案手法）	精度(%)	61.9	—	—	72.0	—	—
	再現率(%)	14.7	—	—	5.6	—	—

表 5 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
Word（比較手法）	精度(%)	68.4	55.3	50.7	75.1	49.8	66.4
	再現率(%)	60.6	37.0	20.8	67.4	19.2	48.7
IG（提案手法）	精度(%)	66.7	60.2	54.9	77.2	58.9	65.9
	再現率(%)	64.0	33.7	31.8	69.9	34.3	51.0

表 6 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
Word（比較手法）	精度(%)	72.7	56.6	77.5	72.2	82.0	70.9
	再現率(%)	70.6	43.3	41.7	69.6	65.4	61.1
IG（提案手法）	精度(%)	71.1	81.4	90.1	71.8	90.8	80.7
	再現率(%)	44.4	11.9	33.3	47.5	20.4	32.4

語を素性に使用する IG 手法の有効性を示すことができた．

しかし，IG 手法による旅行ブログエントリのタイプ分類の精度向上は，旅行ガイドブックのページや質問応答コンテンツに比べ小さい値であった．その原因について考察を行う．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単語の異なり語の割合を，以下の式(1)により求め，結果を表 7 に示す．

$$\frac{\text{あるコンテンツに出現した単語の異なり数}}{\text{あるコンテンツにおいて出現した単語の延べ数}} \quad (1)$$

表 7 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異なり語の割合

	異なり語の割合
旅行ガイドブック	0.082
旅行ブログエントリ	0.112
質問応答コンテンツ	0.078

表 7 より，旅行ブログエントリは，旅行ガイドブックと質問応答コンテンツに比べ，異なり語の割合が高いことがわかる．異なり語の割合が高いと，訓練用データに出現しない単語が，評価用データに出現する場合が多くなる．また，旅行ブログエントリでは，人手でタイプ「買う」や「泊まる」に分類された件数が少なく，訓練用に使用できるデータが少ない．そのため，訓練用データから情報利得により手掛かり語を収集する手法では，手掛かり語を網羅的に収集することができず，タイプ分類の精度を改善することができなかったと考えられる．また，旅行ブログエントリでは，旅行ガイドブックや質問応答コンテンツに比べ，くだけた表現が多用される事も，精度低下の一因であると考えられる．

旅行ブログエントリの各タイプの実験結果について考察を行う．旅行ブログエントリの各タイプにおいても，単語の異なり語の割合を，式(1)を利用して求め，図 5 に示す．図 5 より，各タイプにおいても，異なり語の割合が高いと，タイプ分類の精度が低下する傾向があることが分かった．

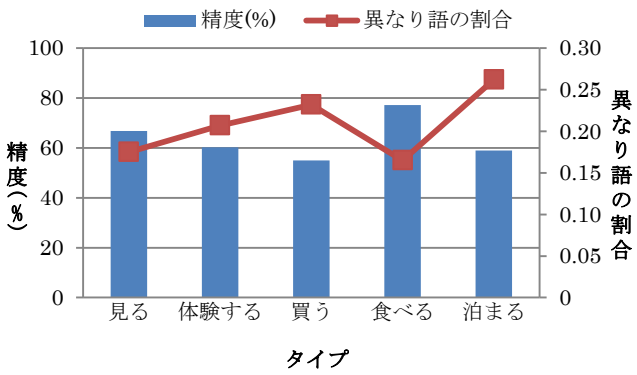

図 5 旅行ブログエントリのタイプ分類の精度と異なり語の割合

旅行ブログエントリのタイプ分類において，最も精度が低かったタイプ「買う」について考察を行う．本研究では，人手によりタイプ分類の正解判定を行う際に，1件の旅行ブログエントリに対して，複数のタイプに判定することを許している．そこで，タイプ「買う」と判定された旅行ブログエントリが，他のタイプに判定された割合を求めた．結果を表 8 に示す．

表 8 タイプ「買う」と判定された旅行ブログエントリが他のタイプに判定された割合

タイプ	他のタイプに判定された割合
見る	0.39
体験する	0.15
買う	0.28
食べる	0.50
泊まる	0.09

表 8 により，人手によりタイプ「買う」と判定されている旅行ブログエントリは，タイプ「食べる」やタイプ「見る」にも判定されている割合が高い．そのため，情報利得により手掛かり語を収集する際に，タイプ「食べる」やタイプ「見る」に関する手掛かり語も収集されてしまう可能性が高いと考えられる．そのため，タイプ「買う」では，タイプ分類の精度向上が小さかったと考えられる．今後は，本研究で構築したタイプ分類のモデルを使用し，多くの旅行ブログエントリのタイプ判定を行うことで，タイプ「買う」のみに判定された旅行ブログエントリを用いて情報利得により手掛かり語の収集を行うことで，さらなる精度の向上が期待できると考えられる．

次に，旅行ガイドブックのページのタイプ分類での，提案手法である 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 61.9%，再現率 14.7%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 0.05 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 9 にまとめる．表 9 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 9 タイプ「見る」における文字数ごとの実験結果

文字数	旅行ガイドブック(件数)	IG 手法		IG+BoVW 手法	
		精度(%)	再現率(%)	精度(%)	再現率(%)
100	151	0.0	0.0	67.9	3.4
200	32	62.7	18.1	66.7	20.0
300	24	25.0	9.4	28.6	12.5
400	34	56.3	29.4	58.5	31.1
500	26	41.7	24.0	41.7	24.0

表 9 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．

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研究として，山口ら[山口 09]の研究がある．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し，画像領域のみから Bag of Visual Words を作成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．

なお，IG 手法において，文字数ごとに精度のばらつきがあるのは，旅行ガイドブックから OCR 処理を行うことでテキスト情報を抽出しているためであると考えられる．テキスト情報には，OCR 処理を行う際に文字化けが多く発生しており，テキストの文字数が，タイプ分類を行う際に有益なテキストの情報の量に比例しないためだと考えられる．

5.2 旅行ガイドブックのブック単位への対応付け

§実験条件

実験に用いるデータと評価尺度

実験には，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 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，以下に示す提案手法と，4 種類の比較手法について実験を行った．4.2 節では，k 近傍法を使用した手法（KNN 手法）を提案したが，比較手法として，機械学習を使用した手法（SVM 手法）を使用する．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．

KNN_TYPE 手法においてのみタイプ分類の結果を使用する．Step 1 により，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は判定済である．実験では，タイプは，旅行ガイドブックのページは IG+BoVW 手法，旅行ブログエントリと質問応答コンテンツは IG 手法により自動で判定された結果を使用する．KNN_TYPE 以外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，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から抽出した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．

< 比較手法 >

- BASE_KNN : KNN 手法を用いる．類似度の計算には，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から抽出した名詞の出現頻度を使用する．
- BASE_SVM : SVM 手法を用いる．素性には，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から抽出した名詞の出現頻度を使用する．
- KNN_LOC : KNN 手法を用いる．類似度の計算には，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から抽出した地名の出現頻度を使用する．
- SVM_LOC : SVM 手法を用いる．素性には，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から抽出した地名の出現頻度を使用する．

§ 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 10, 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 11 に示す．

表 10 旅行ガイドブックと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	27.3	15.5
	BASE_SVM	21.6	3.0
	KNN_LOC	76.7	20.1
	SVM_LOC	44.0	19.0
提案手法	KNN_TYPE	81.1	20.4

表 11 旅行ガイドブックと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	48.7	16.6
	BASE_SVM	40.5	18.4
	KNN_LOC	78.3	20.6
	SVM_LOC	39.8	30.1
提案手法	KNN_TYPE	85.8	21.0

表 10, 表 11 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. 本研究では，誤った情報が旅行ガイドブックに対応付けられるよりも，適切な情報が対応付けられたほうが，システムとして有益であると考え, 再現率よりも精度を重要視する. また，本研究では，地名とタイプ分類を利用した対応付け手法を採用しており，すべての出現単語を使用した対応付けに比べ柔軟な対応付けが可能である．そのため，再現率が低くても，様々な情報を含んだ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と考えている．

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．本研究では，各コンテンツから日本語構文解析器 CaboCha を使用することで地名を抽出し，対応付けを行っているため，旅行先ではない地名が抽出されると，判定を誤ってしまう．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[Ishino 11b]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．

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から地名のみを抽出し，対応付けを行ったが，地名以外にも，土産物や特産物の名前など，その地域を連想させる単語を抽出できれば，より正確に対応付けを行うことができると考えられる．地域を連想させる単語を収集する研究がある[晃昇 13, 奥 12]．これらの研究により収集された地域を連想させる単語を利用することで，更なる精度，再現率の向上が可能であると考えられる．

5.3 旅行ガイドブックのページ単位への対応付け

§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

提案手法の有効性を確かめるため，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，Step 2 の旅行ガイドブックのブック単位の対応付けにより，対応付ける旅行ガイドブックが判定済みである．実験には，5.2 節の KNN_TYPE 手法により自動で対応付けられた旅行ガイドブックを使用する．また，5.2 節の実験と同様に，タイプは，旅行ガイドブックのページは IG+BoVW 手法，旅行ブログエントリと質問応答コンテンツは IG 手法により自動で判定された結果を使用する．

比較手法では，旅行ガイドブックのブック単位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 >

- 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．

§実験結果と考察

表 12 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 12 対応付け結果が「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	0.72	0.53
提案手法 1	0.82	0.57
提案手法 2	0.82	0.77

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．構築したシステムの有用性を確かめるため，有用性評価 1 と有用性評価 2 を行った．

有用性評価 1

本研究で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 6 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 6 より，旅行ブログエントリ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．また，被験者による自由記述を表 13 に示す．

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対応箇所を強調して表示したり，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[Yasuda 06, Ikeda 08, Schler 06]．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．

有用性評価 2

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自動で対応付ける手法を提案している．また，提案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．構築したシステムの有用性を確かめるため，被験者 3 名に対し，構築したシステムを使用する場合と，旅行ガイドブックを閲覧し，人手で旅行ガイドブックのページに適切に対応付く旅行ブログエントリと質問応答コンテンツを検索する場合とで比較を行った．まず，被験者 3 名に，50 ページのガイドブックを閲覧させ，各ページに適切に対応付く旅行ブログエントリと質問応答コンテンツを 1 件ずつ発見するまでの時間を計測した．旅行ブログエントリは Google ブログ検索[14]，質問応答コンテンツは Yahoo!知恵袋の検索システム[15]（カテゴリは「地域，旅行，お出かけ」に限定）を使用し，デスクトップパソコンを使用して検索することとした．その結果，旅行ガイドブックのページ 50 件に対し，被験者 3 名の平均で旅行ブログエントリでは約 79 分，質問応答コンテンツでは約 47 分かかることがわかった．旅行ガイドブックは 1 冊平均 145 ページであるが，1 冊の旅行ガイドブックの各ページに 1 件ずつ旅行ブログエントリを対応付けるためには 3.5 時間以上，質問応答コンテンツでは 2 時間以上必要となり，高コストを要する．

また，被験者に対し，構築したシステムを利用する場

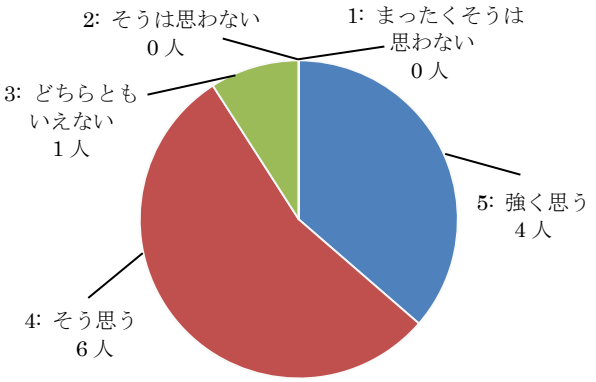

図 6 情報拡張された旅行ガイドブックを閲覧するシステムの有用性評価

表 13 有用性評価 1 における被験者による自由記述の結果

利点／欠点	自由記述
利点	・ 旅行ガイドブックだけでは得られないローカルな情報など，より観光地の詳しい情報を得ることができる． ・ 季節や天候などによってのおすすめポイントや観光ルートなどの情報を得ることができる． ・ 子連れの家族旅行でのおすすめのスポットなど，旅行形態に応じた情報を得ることができる．
欠点	・ 旅行ブログの場合，文章が長いものがあり，読むのに時間がかかる． ・ 食事や，観光スポットへの感想は，人によって感じ方が違うのではないかと思う．

*14 http://www.google.co.jp/blogsearch
*15 http://chiebukuro.search.yahoo.co.jp/advanced?p=&ei=UTF-8&class=1

合と，人手で旅行ガイドブックに適切に対応付ける旅行ブログエントリと質問応答コンテンツを検索する場合とで比較してもらい，使用感についてアンケート調査を行った．被験者による自由記述を表 14 に示す．

表 14 有用性評価 2 における被験者による自由記述の結果

	自由記述
①	検索するキーワードの選定に時間がかかった. 旅行ガイドブック内に掲載されている観光名所やレストランなど, 固有名をキーワードに使用すればその場所に関するピンポイントの情報を得ることができたが, 構築システムから得られるような様々な情報を得ることができなかった.
②	イタリアのレストランを紹介した海外の旅行ガイドブックのページに関連する旅行ガイドブックを検索すると, 日本にあるイタリアンレストランに関連する旅行ブログエントリが大量に検索され, イタリアにあるレストランに関連する旅行ブログエントリを検索するのに時間がかかった.

本研究では，地名とタイプ分類を使用した対応付け手法を提案しシステムを構築しているため，表 13 の有用性評価 1 の自由記述からわかるように，様々な情報を対応付けることができている．しかし，表 14 の①が示すように，実際に検索を行う際には，キーワードの選定に時間かかることがわかった．

また，表 14 の②のように，海外の旅行ガイドブックのページに対応づく旅行ブログエントリや，質問応答コンテンツを検索することは，特に困難であることがわかった．本研究では，まず旅行ガイドブックのブック単位へ旅行ブログエントリや質問応答を対応付けている．そのため，イタリアの旅行ガイドブックに，日本のレストランの情報が付与されることを防ぎ，その旅行ガイドブックの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ことが可能である．

有用性評価 1 と有用性評価 2 の結果より，提案システムは，ユーザが情報を得るためのコストを低下し，様々な情報を含む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に対応付けることが可能であることを示すことができたため，本研究で構築したシステムは有効であるといえる．

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った．旅行ガイドブックのページでは精度 75.8%，再現率 33.7%，旅行ブログエントリでは精度 65.9%，再現率 51.0%，質問応答コンテンツでは精度 80.7%, 再現率 32.4%を得た． Step 2 の旅行ガイドブックのブック単位への対応付けでは，旅行ブログエントリでは精度 81.1%，再現率 20.4%，質問応答コンテンツでは精度 85.8%，再現率 21.0%を得た．Step 3 の旅行ガイドブックのページ単位への対応付けでは，旅行ブログエントリでは 82.2%，質問応答コンテンツでは 77.0%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，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 ◇

[Broder 07] 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).

[Csurka 04] 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).

[Davidov 09] 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).

[群 06] 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).

[藤坂 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).

[He 10] 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).

[Ikeda 08] Ikeda, D., Takamura, H. and Okumura, M.: Semi-supervised Learning for Blog Classification, Proc. 23rd AAAI Conference on Artificial Intelligence, pp.1156-1161 (2008).

[Ishino 11a] 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).

[Ishino 11b] 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).

[神谷 00] 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 欧米 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).

[Kando 97] 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).

[晃昇 13] 晃昇祥恵, 森田和宏, 泓田正雄, 青江順一: 地域連想語辞書の構築に関する研究, 言語処理学会第 18 回年次大会 (2012).
[Nanba 09] 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).
[岡本 10] 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).
[奥 12] 奥健太, 西崎剛司, 服部文夫: 地域限定性スコアに基づく位置情報付きコンテンツからの地域限定語句の抽出, 情報処理学会論文誌 データベース, Vol.5, No.3 (TOD55), pp.97-116 (2012).
[Rakesh 11] 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).
[斉藤 12] 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, 第 5 回 Web とデータベースに関するフォーラム, (2012).
[Salton 71] Salton, G: The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ (1971).
[Schler 06] 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).
[徳久 11] 徳久雅人, 村田真樹: 観光開発のヒントをブログ記事から得るための支援技術～SVM を用いる場合～, 第 8 回観光情報学会全国大会発表概要集, pp.44-45, (2011).
[山口 09] 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).
[柳井 07] 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).
[Yang 07] 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).
[Yasuda 06] 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).
[渡邊 11] 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, (2011).

〔担当委員：岡崎 直観〕

2014 年 1 月 8 日　受理

著 者 紹 介

石野　亜耶
2009 年広島市立大学情報科学部知能情報システム工学科卒業．2011 年広島市立大学大学院情報科学研究科博士前期課程修了．2014 年，同大学大学院情報科学研究科博士後期課程満期退学．同年広島経済大学経済学部ビジネス情報学科助教．現在に至る．言語処理学会，情報社会学会各会員

藤井　一輝
2013 年広島市立大学情報科学部知能工学科卒業．現在，広島市立大学大学院情報科学研究科博士前期課程在学中．

藤原　泰士
2013 年広島市立大学情報科学部知能工学科卒業．現在，広島市立大学大学院情報科学研究科博士前期課程在学中．

前田　剛
2013 年広島市立大学情報科学部知能工学科卒業．現在，広島市立大学大学院情報科学研究科博士前期課程在学中．

難波　英嗣（正会員）
1996 年東京理科大学理工学部電気工学科卒業．1998 年北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．2001 年同大学情報科学研究科博士後期課程修了．同年日本学術振興会特別研究員．2002 年東京工業大学精密工学研究科助手．同年広島市立大学情報科学部講師．2010 年広島市立大学大学院情報科学研究科准教授．現在に至る．博士（情報科学）．テキストマイニング，情報検索，自動要約，特許情報処理に関する研究に従事．言語処理学会，情報処理学会，地理情報システム学会，ACL，ACM 各会員．

竹澤　寿幸（正会員）
1984 年早稲田大学理工学部電気工学科卒業．1989 年同大学大学院理工学研究科博士後期課程修了．工学博士．同年（株）国際電気通信基礎技術研究所入社．2007 年広島市立大学大学院情報科学研究科教授．現在に至る．音声対話翻訳，感性コミュニケーションの研究開発に従事．2006 年電子情報通信学会 ISS 論文賞受賞．電子情報通信学会，情報処理学会，日本音響学会，言語処理学会各会員．