広島市立大学審査博士学位論文

ソーシャルメディアを利用した
観光支援システムの自動構築

2014 年 9 月

石野 亜耶

要旨

2007 年 1 月に「観光立国推進基本法」が施行され，2008 年 10 月には国土交通省の外局として観光庁が設置されるなど，日本では「観光」を 21 世紀の基幹産業と位置付け，観光を支援するための多様な取り組みが積極的に推進されている．観光を支援する媒体としては，地方公共団体や旅行会社などが運営する観光ポータルサイトや，「るるぶ」などに代表される旅行ガイドブックが挙げられる．

観光ポータルサイトや旅行ガイドブックでは，土産物や観光名所などの情報，ホテルやレストランへのリンクなどの基本的な観光情報が紹介されている．しかし観光情報は，土産物の商品開発や，テーマパークなどの施設の建設などにより日々新しくなり，ホテルやレストランを紹介する Web ページも新しく作成される．そのため観光情報を新たに獲得し，古くなった情報は削除するといった更新作業が不可欠である．しかし，既存の観光情報データベースは，人手で観光情報を収集し，整理，保守するため，非常に時間とコストがかかるといった問題点がある．

近年，ソーシャルメディアの普及により，個人からの情報発信が盛んになってきている．ソーシャルメディアの例として，ブログ，Twitter，質問応答コンテンツが挙げられる．旅行者が気軽に観光情報を発信する場として，ブログが使用されている．旅行記が記述されたブログエントリを旅行ブログエントリと呼ぶこととする．旅行ブログエントリには，土産物，観光名所，旅行の際に参考にした Web ページへのリンクなど，過去の旅行者が体験した様々な観光情報を含んでいる．そこで本研究では，観光情報源の有益な情報源として旅行ブログエントリに着目し，手掛かり語を用いて，旅行ブログエントリを自動で抽出する手法を提案した．また実験により提案手法の有効性を示した．旅行ブログエントリの検出に関しては，精度 86.7%，再現率 38.1%を得た．また，旅行ブログエントリから，土産物情報，観光名所情報を抽出することで，旅行ブログエントリの観光情報の情報源としての有用性を確認した．

本研究では，自動で抽出した旅行ブログエントリを利用し，2 種類の観光支援システムを構築した．自動で抽出した旅行ブログエントリを利用することで，低コストで観光支援システムを作成することが可能になると考えられる．同時に，網羅性の高さや最新の観光情報を素早く獲得できる点などで，既存の観光を支援する媒体よりも有用なものになることが期待される．

旅行ブログエントリ中に含まれる観光情報を利用した観光支援システムとして，観光情報リンク集を構築した．旅行ブログエントリには，観光の際に参考にした Web ページへのリンクが，観光情報として提示されている．旅行ブログエントリ中に含まれるリンクを，

観光情報リンクと呼ぶこととする．本研究では，旅行ブログエントリから観光情報リンクを収集し，「見る」，「食べる」などの観光に特化したタイプに分類することで，自動で観光情報リンク集を構築した．タイプ分類には，観光情報リンク周辺の文字列を使用した．実験の結果，精度 76.9%，再現率 66.4%を得た．

また，ソーシャルメディア上の投稿を，既存の観光情報データベースと組み合わせた観光支援システムとして，情報拡張した旅行ガイドブックの閲覧システムを構築した．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．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のか，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，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と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手法を提案する．実験の結果，旅行ガイドブックのページへ，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，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できるシステムの構築を行い，被験者による評価により，提案システムが旅行の計画を行う際に有用であることを示した．

本研究では，日本語で記述された旅行ブログエントリを対象としたが，英語で記述された旅行ブログエントリも対象とすることで，外国人旅行客の観光支援システムの構築への応用も可能である．また，本研究では主に，旅行ブログエントリに含まれるテキスト情報を使用して観光支援システムを構築している．今後の研究では，flickr などの画像共有サイトや YouTube などの動画共有サイトに投稿された写真や動画も対象とすることで，より視覚的な情報も提供できる観光支援システムを構築していく予定である．

キーワード: 観光情報処理，観光支援システム，ソーシャルメディア，旅行ブログエントリ

Automatic Construction of Sightseeing Support Systems using Social Media

Aya Ishino

In Japan, the Basic Act for Promoting a Tourism-oriented Country was put into force in January 2007. The Japan Tourism Agency was launched in October 2008, and aims to promote domestic and overseas tourism. Travelers planning to visit particular tourist spots need information about these tourist destinations, and they often use travel guidebooks to collect this information. Guidebooks give basic information about famous tourist spots, souvenir shops, and restaurants. Besides, there are many portal sites that are operated by travel companies and local governments for the benefit of the tourist. However, it is costly and time-consuming to compile travel information for all tourist spots and to keep this data up to date manually.

Recently, with explosive spread of social media, users can express their ideas and opinions on the Internet easily and actively. Examples of such social media include blogs, micro blogs, and archives of questions answered (QA archives). Many travelers have been writing their travel experiences via blogs, or "travel blog entries". For example, some bloggers introduce web sites about tourist spots, while others report on location names and local products.

To compile such travel information automatically, we focus on travel blog entries.

There are portal sites that have travel blog entries, such as "TravelBlog" (http://www.travelblog.org) and "4travel" (http://4travel.jp). These portal sites publish travel blog entries collected from user's blog entries. However, not all travel blog entries are registered on the portal sites. Aiming to construct an exhaustive database of travel blog entries, we have studied the automatic identification of travel blog entries from a blog database. Basically, travel blog entries that contain cue phrases, such as "旅行" (travel), "観光" (sightseeing), or "ツアー" (tour), have a high degree of probability of being travel blog entries. However, not every travel blog entry contains such cue phrases. Therefore, we formulate the identification task of travel blog entries as a sequence-labeling problem, and employ machine learning techniques to solve it. We conducted experiments on travel blog identification and obtained scores of 86.7% for Precision and 38.1% for Recall. To investigate the effectiveness of travel blog entries as a source for travel information, we extracted souvenir information and tourist spots information as travel information from them. For our experiments on travel information extraction from travel blog entries, we obtained 74.0% and 71.0% for Precisions at the top 100 extracted local products and tourist spots, respectively, thereby confirmed that travel blog entries are a useful source of travel information.

In this paper, we constructed two systems for sightseeing support using travel blog entries. We expect our method with automatically extracted travel blog entries to be helpful for constructing sightseeing systems at low cost. Furthermore, our method has an advantage over conventional systems in that it takes less time for the latest travel information to be reflected in the system.

One is a system about collection of travel information links. To collect the information about tourist spots, travelers use portal sites. However, as there are also sites that are not updated frequently and there are large differences in the amount of information on each tourist destination. Some bloggers introduce useful web sites for a tourist spot, an accommodation and a restaurant in travel blog entries. Here, we extracted hyperlinks

from travel blog entries and constructed the collection of travel information links. Our method consists of two-steps. First, we extract a hyperlink and any surrounding sentences that mention the link (a citing area). We manually created rules for the automatic extraction of citing areas. These rules use cue phrases. When bloggers of travel blog entries introduce web sites, quotation marks or brackets are often used immediately before and after the title of the site. The bloggers also use particular words, such as “紹介” (introduction), “公式サイト” (official site), or “の HP” (web page of), or particular marks, such as quotation marks or brackets. Therefore, we manually selected 26 cues and used them for citing area extraction. Second, we classify the link by taking account of the information in the citing area. We classify the link into four types of content, such as “Spot” and “Hotel.” We employed a machine-learning technique. For classification of the link type, we use cue phrase. For example, for classification of type “Spot”, we use words used in the names of tourist spots, such as “動物園” (zoo) or “博物館” (museum). We employed a machine-learning technique. For classification of link types, we obtained scores of 76.9% for Precision and 66.4% for Recall. Finally, we have constructed a system that can search for travel information links. If user inputs a keyword in the search form and clicks the “link” button, the system shows a list of links for web sites related to the keyword together with automatically identified link types and citing areas.

The other is a system that provides enriched guidebooks using travel blog entries and QA archives. Travelers planning to visit a particular tourist spot need information about their destination and they often use travel guidebooks to collect this information. Guidebooks give basic information about famous tourist spots, souvenir shops, and restaurants. Travelers only get basic information from them. However, they cannot determine how to travel between the destinations listed in the guidebooks or at which hotel to stay with their family. To help travelers read guidebooks, we focused on travel blog entries and QA archives, both of which contain valuable information, such as first-hand accounts by users who have visited the particular destination. In addition, blog entries and QA archives have more descriptive content than other social media services, such as twitter or various review sites. Therefore, travel blog entries and QA archives are considered useful information sources for obtaining travel information.

Our method consists of three steps. First, we classify pages of guidebooks, travel blog entries and QA archives into five types of content, such as “watch” and “eat.” For classification them, we collected cue phrase that were peculiar to each content type for machine learning, and employed information gain as the feature selection method. Guidebooks contain many images. Therefore, for the content-type classification of pages of guidebooks, we also employed image information as features for machine learning. We adopted the Bag of Visual Words for image information. Second, we align each travel blog entry and QA archive with guidebooks by taking these content types into account. Third, we align each travel 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 travel 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. By using our system, users can obtain basic information from guidebooks and valuable information based on the personal experiences of travelers from travel blog entries and QA archives.

In this work, we focused on travel blog entries written in Japanese. Our future work

involves translating cue phrases from Japanese into other languages, and applies our method into sightseeing support systems in various languages. In this paper, we focus on text information on travel blog entries and QA archive as social media. We will also research on using images and movies posted at hosting websites such as flickr and YouTube for our systems to provide visual information.

Keyword: travel information processing, sightseeing support system, social media, travel blog entry

目次

図目次

表目次

第1章　序論

Innsbruck, Austria, 2011

本章では，本研究の背景と目的について述べる．1.1 節では，研究の背景，1.2 節では研究の目的について説明を行う．1.3 節では，本論文の構成について述べる．

1.1 研究の背景

2007 年 1 月に「観光立国推進基本法」が施行され，2008 年 10 月には国土交通省の外局として「観光庁」が設置されるなど，日本では「観光」を 21 世紀の基幹産業と位置付け，観光を支援する多様な取り組みが積極的に推進されている*1.

旅行を計画する際には，旅先の観光名所，土産物，飲食店，宿泊施設などの観光情報を収集する．これらの観光情報を収集する情報源を提供し，観光を支援する媒体としては，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，地方公共団体や旅行会社などが運営する観光ポータルサイトが挙げられる．しかし，上記の様な既存の観光情報データベースは人手で観光情報を抽出し，整理，保守するため，非常に時間とコストがかかるといった問題点が挙げられる．

そこで，本研究では，旅行者が気軽に観光情報を発信する場としてよく用いられるブログに注目した．本研究では，旅行記が記述されたブログエントリを，旅行ブログエントリと呼ぶ．旅行ブログエントリには，土産物，観光名所，旅行の際に参考にした Web ページへのリンクなど，過去の旅行者が体験した様々な観光情報を含んでいる．このような旅行ブログエントリから自動的に観光情報を抽出することで，低コストで観光支援システムを作成することが可能になると考えられる．同時に，網羅性の高さや最新の観光情報を素早く獲得できる点などで，既存の観光を支援する媒体よりも有用なものになることが期待される．本研究では，旅行ブログエントリを利用し，2 種類の観光支援システムを構築する．

1.2 研究の目的

本論文では，観光情報の情報源として，ブログデータベースから旅行ブログエントリを自動的に抽出する手法を提案する．自動で抽出した旅行ブログエントリから，観光情報として，土産物情報と観光名所情報を自動で抽出することで，観光情報の情報源としての有用性評価を行う．また，旅行ブログエントリに含まれるリンクを利用することで，観光情報リンク集を自動構築する．また，自動で抽出した旅行ブログエントリを使用し，観光データベースを横断して提示するシステムの一例として，旅行ガイドブックの情報拡張を行う．

*1 国土交通省 観光庁　施策　http://www.mlit.go.jp/kankocho/shisaku/index.html

1.2.1　旅行ブログエントリの自動抽出および有用性評価

本研究では，観光情報を抽出するための情報源として，旅行ブログエントリに注目する．旅行ブログエントリには，土産物，観光名所，旅行の際に参考にしたWebページへのリンクなど，過去の旅行者が体験した様々な観光情報を含んでいる．このような旅行ブログエントリから自動的に観光情報を抽出することで，低コストで観光情報データベースを作成することが可能になると考えられる．同時に，網羅性の高さや最新の観光情報を素早く獲得できる点などで，既存の観光ポータルサイトよりも有用なものになることが期待される．

そこで本研究では，観光情報源の有益な情報源として旅行ブログエントリの抽出を試みる．本研究では，旅行ブログエントリに含まれるテキスト情報を使用することで，ブログデータベースから旅行ブログエントリを自動で抽出する手法を提案し，実験により提案手法の有効性を確認する．

また，旅行ブログエントリは，観光情報の抽出のための有用な情報源であることを確かめるため，旅行ブログエントリから観光情報を自動で抽出する．本研究では，土産物，観光名所に関する情報を観光情報として抽出する．また，一般のブログエントリ，Web 文書からも観光名所情報・土産物情報を試みることにより，旅行ブログエントリの観光情報の情報源としての有用性評価も行う．

1.2.2　観光情報リンク集の自動構築

旅行ブログエントリ中に含まれる観光情報を利用した観光支援システムとして，観光情報リンク集を構築した．Web 上で利用可能な観光を支援する媒体の一例として，地方公共団体や旅行会社などが運営する観光ポータルサイトが挙げられる．観光ポータルサイトでは，ホテルやレストランへのリンクが観光情報として紹介されている．しかしホテルやレストランを紹介するWebページも新しく作成されるため，観光情報を新たに獲得し，古くなった情報は削除するといった更新作業が不可欠である．しかし，既存のデータベースは人手で観光情報を抽出し，整理，保守するため，非常に時間とコストがかかる．

旅行ブログエントリには，観光の際に参考にしたWebページへのリンクが観光情報として提示されている．旅行ブログエントリ中に含まれるリンクを，観光情報リンクと呼ぶこととする．そこで，旅行ブログエントリから自動的に観光情報リンクを収集し，「食べる」や「見る」などの観光に特化したタイプへ分類することで，低コストでの観光情報リンク集を構築する．同時に，網羅性の高さや最新の観光情報を素早く獲得できる点などで，既存の観光ポータルサイトよりも有用なものになることが期待される．

1.2.3　旅行ガイドブックの情報拡張

ソーシャルメディア上の投稿を，既存の観光情報データベースと組み合わせた観光支援システムとして，情報拡張した旅行ガイドブックの閲覧システムを構築した．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．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のか，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，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と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手法を提案する．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．

1.3　論文の構成

2 章では，観光情報処理に関する諸研究やサービスをサーベイし，既存の研究との違いを明らかにする．

3 章では，ブログデータベースから，旅行ブログエントリを自動で抽出する手法を説明する．旅行ブログエントリの自動抽出手法の有効性を確認するため，実験を行う．また，旅行ブログエントリから観光情報として，土産物情報と観光名所情報を抽出することで，旅行ブログの観光情報の情報源としての有用性評価を行う．

4 章と 5 章では, 旅行ブログエントリを使用した観光情報の抽出システムについて説明する．

4 章では，旅行ブログエントリを利用した観光情報リンク集の自動構築について述べる．提案手法の有効性を確認するための実験を行う．また，観光情報リンク集の検索システムを構築し，実際のシステムの動作例を示す．

5 章では, 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

拡張を行う手法について述べる．実験により提案手法の有効性を確認する．また，提案手法を用い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，その動作例を示す．

6 章では，結論と今後の課題について述べる．

第2章　観光情報データベース

Chiang Mai, Thailand, 2011

本章では，観光情報データベースに関連する諸研究やサービスを紹介し，本研究との違いを明らかにする．2.1 節では，既存の観光情報データベースを紹介する．2.2 節では，Web を利用した観光情報データベースの自動構築，2.3 節ではソーシャルメディアを利用した観光情報データベースの自動構築，2.4 節では，地理情報検索についてについて説明する．

2.1　既存の観光情報データベース

観光を支援する媒体としては，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，地方公共団体や旅行会社などが運営する観光ポータルサイトが挙げられる．図 2.1 は，一般社団法人広島県観光連盟により運営されている広島県の観光サイト「ひろしま観光ナビ」*2の「観光スポット」のページである．このページでは，観光スポットとして，観光名所や飲食店の紹介やホームページへのリンクが紹介されている．

http://www.kankou.pref.hiroshima.jp/spot/index.html

図 2.1　「ひろしま観光ナビ」の「観光スポット」のページ

*2 ひろしま観光ナビ　http://www.kankou.pref.hiroshima.jp/index.html

しかし，上記の様な既存の観光情報データベースは人手で観光情報を収集し，整理，保守するため，非常に時間とコストがかかるといった問題点が挙げられる．この問題を解決するために，Web を利用することで観光情報データベースを自動構築する試みがある．

2.2 W eb を利用した観光情報データベースの自動構築

Web から地域情報を自動収集しようとする研究がある[1]．大槻らは，地方公共団体ドメイン名に対応する URL を作成し，これらの URL とのリンク関係を利用することで，日本の全地域(3427 自治体)の 80%以上に対して，情報源となる地域サイトを収集することに成功している．

佐藤[2]は，Web を利用した住所探索を提案している．この提案手法では，まず検索エンジンを利用して住所情報が記載されている可能性が高い Web ページを収集し，その Web ページから住所データを抽出する．このようにして得られた住所データを整理・統合して，目的の住所情報を出力する．

村山ら[3]は，位置情報が明記されていない文書に位置情報をメタデータとして付与するため，お店の名前等の固有名と対応する位置情報のデータベースを Web 上の文書から自動的に作成する手法を提案している．

Davidov[4]は，Web から交通手段や経路の地理的なネットワークを見つける手法を提案している．

観光情報の中でも，祭りやイルミネーション，マラソン大会などのイベントに関する情報は，旅行の計画を立てる観光客にとって重要な情報である．例えば，今から訪れようとしている場所で何が行われるのか，それがどんな様子なのか，といった情報が分かれば，旅先での行動が計画しやすくなる．斉藤ら[5]は，新聞記事からイベント情報を抽出する手法を提案している．現在より未来に開催される娯楽に関する催しや行事が記載されている新聞記事を「イベント記事」とし，新聞記事がイベント記事かを自動検出する．そして検出したイベント記事より，手がかり語の有無を素性とした機械学習を用いてイベント名，開催日時，開催地，開催施設名といったイベント情報を抽出している．斉藤らにより構築されたイベント検索システムを図 2.2 に示す．

((斉藤 12)[5]より抜粋)

図 2.2　イベント情報検索システム

2.3　ソーシャルメディアを利用した観光情報データベースの自動構築

近年，ソーシャルメディアの発達に伴い，Web 上でも特に，ソーシャルメディア上の投稿を，観光情報データベースの構築に利用する研究がある．

本研究と同様に，ブログを情報源とし，観光情報を自動抽出する研究がある[6, 7]．岡本ら[6]は，一般のブログ検索エンジンを利用することで，地名を含むブログエントリを収集し，それらのブログエントリから，地域イベント情報を抽出する手法を提案している．徳久ら[7]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．

ブログからユーザの行動経路を抽出する研究がある[8, 9]．郡ら[8]は，ブログからユーザの行動時の代表的な経路とその文脈を抽出し，それらを地図上にマッピングすることにより，集約して提示するシステムを提案している．Kori ら[9]は，ローカルブログエントリから典型的な旅行ルートを抽出し，これらのルートに沿って関連するマルチメディアコンテンツを提示するシステムを提案している．

しかし，ブログエントリの中には観光と関係ないものも存在するため，ブログエントリを全て使うと，十分な精度で観光情報が抽出できない可能性がある．本研究では，まず，ブログ集合から旅行ブログエントリを自動検出し，次に，そこから観光情報の抽出をすることにより，高い精度での抽出を目指す．

Twitter に代表されるマイクロブログから観光情報を抽出する研究がある．Twitter からイベント情報を抽出する研究を紹介する．藤坂ら[10]は，Twitter から地域イベントを発見し，その特性を検証するためのシステムを提案している．ジオタグ付きツイートを利用し，領域ごとのツイート数を，通常時のツイート数と比較することで，イベント情報を抽出している．金子ら[11]は，Twitter からイベント情報を抽出する他に，ツイートから抽出したキーワードを用いて，画像を収集し，代表画像を地図上にマッピングすることで，イベントの様子を視覚的に捉えやすくしている．奥ら[12]は，ツイートや写真共有サイトに投稿されたジオタグ付きのデータを利用することで，観光スポットの特徴を抽出し，観光スポットの推薦システムを構築している．

画像共有サイトや動画共有サイトに投稿された画像や動画に着目した研究がある．Arase ら[13]は，画像共有サイトに投稿された写真のメタデータ(ジオタグ[緯度，経度]，写真に関する簡単な説明文・単語)を利用してユーザの旅行パターンのマイニングを行っている．Rattenbury ら[14]は，画像共有サイトに投稿された画像に付与されたメタデータを使用し，タグから地名やイベント名の抽出を行っている．

島田ら[15]は，本研究では，イベント参加支援システムを構築することを目標にイベントに関する動画の自動収集を行った．島田らにより構築されたイベント検索システムを図 2.3 に示す．地図に表示されたピンをクリックすると，イベント動画を閲覧することができる．

((島田 15)[15]より抜粋)

図 2.3 イベント動画検索システム

2.4 地理情報処理

観光情報は，地理情報の一種とみなすことができる．近年，地理情報検索に関する様々な研究が行われている。CLEF(Cross Language Evaluation Forum)という評価ワークショップのタスクのひとつとして，地理系に特化した情報を検索する Geo CLEF*[3]が 2005 年から開催されている[16]．このタスクの目的は，新聞記事集合から「ヨーロッパにある川の周りはワイン作りが盛んな地域だ」のような地理情報の関連記事を探すというものである．本研究では，新聞記事の代わりに，一般の旅行者が気軽に観光情報を発信する場としてよく使うブログに焦点をあてた．

地理情報を検索する際には，対象となる Web ページがどの地域を対象に記述されたのかを正確に判定する必要がある．この様な Web ページの地理的スコープを自動的に抽出する研究がある[17, 18, 19]．Amitty ら[17]は，Web ページに含まれる地名を抽出し，地名の階層関係を利用することで，地理的スコープを求めている．

安田ら[20]は，携帯端末等を持ちながら歩いている状況で，できるだけ近くの店舗に関する情報を検索するといった，小さな領域における距離を考慮した検索を可能とする手法を提案している．この手法は，例えば「東京都新宿区歌舞伎町」周辺の情報を知りたいという検索が行われた場合，「東京都新宿区歌舞伎町」に言及する文書を優先的に検索することができる．

Web から地理情報を抽出する手法の 1 つに，あらかじめ検索対象のリストを作成し，クローリングによって得られた情報を各検索対象に関連づける登録型検索手法がある．しかし，この手法はリストに登録されていない対象に関する情報を抽出できないという欠点がある．そこで相良ら[21]は，実世界に存在する店舗を対象に新規店舗を検索し，店舗データを新たに登録する手法を提案している．この手法は，既に登録されている店舗情報を収集する際に得られた Web ページ群から検索することで，新規店舗候補の検索を効率良く行う．

[3] http://ir.shef.ac.uk/geoclef/

第3章　旅行ブログエントリの自動抽出および有用性の評価

Dubai, United Arab Emirates, 2011

本章では，ブログデータベースから旅行ブログエントリを自動で抽出する手法について述べ，旅行ブログエントリから観光情報を抽出することで情報源としての有用性評価を行う．まず，3.1 節で旅行ブログエントリに関連する研究を紹介する．3.2 節で旅行ブログエントリを定義し，3.3 節では旅行ブログエントリの抽出手法と実験結果について述べる．また，3.4 節では旅行ブログからの観光情報の抽出手法の説明および実験と有用性評価を行う．3.5 節でまとめを行う．

3.1　旅行ブログエントリに関する研究

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” *4，“旅行・観光ブログ村” *5，“フォートラベル” *6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．一般ブログに焦点を当てることで，様々な層のより多くの旅行ブログエントリを収集できると考えられる．そのため本研究では，ブログデータベースから旅行ブログエントリを自動で抽出する手法を提案する．

Ishino ら[22]は，広島の特徴のひとつである広島電鉄の電車（広電）を使用した観光を支援するための枠組みの一つとして，広電の電停に関する旅行ブログ（電停ブログ）を収集している．藤井ら[23]は，収集した電停ブログを地図にマッピングし提示するシステム“ぶらり広島電停散歩 MAP”を構築している．構築されたシステムを図 3.1 に示す．地図上のピンをクリックすると電停ブログを検索することができる．Ishino らは電停ブログを抽出しているが，本研究では，旅行ブログエントリを抽出する．

（(藤井 14)[23]より抜粋 ）

図 3.1　ぶらり広島電停散歩 MAP

*4 http://www.travelblog.org/
*5 http://travel.blogmura.com/
*6 http://4travel.jp/

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[24, 25, 26]．このような技術を利用し，ブログ著者の属性と，観光情報の利用者の属性を照らし合わせることで，例えば「女性に人気の土産物」や「若い人に人気の観光名所」など，利用者に適した観光情報を推薦することができると考えられる。

3.2 旅行ブログエントリの判定基準

本研究では，観光情報が記述されたブログエントリを，旅行ブログエントリであると判定する．観光の定義とは，1995 年に観光政策審議会によって定義された「余暇時間の中で，日常生活圏を離れて行う様々な活動であって，触れ合い，学び，遊ぶということを目的とするもの」*7とする．観光の定義を図 3.2 に示す．観光の定義に従い，旅行準備中の記事や観光情報の含まれない出張・研修・留学中のブログエントリは，旅行ブログエントリではないと判定する．

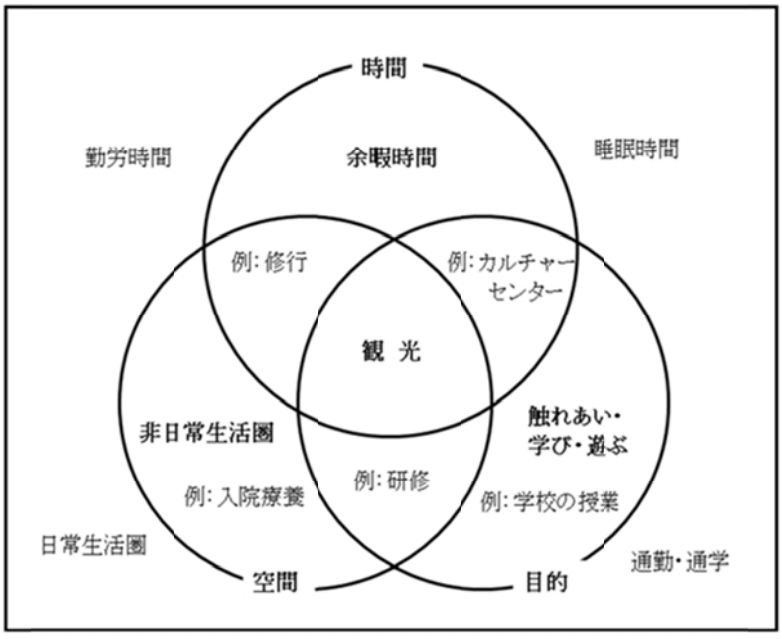


図 3.2 観光の定義

*7 http://www.mlit.go.jp/singikai/unyusingikai/kankosin/kankosin39.html

この判定基準をもとに旅行ブログエントリと判定したブログエントリの例を図 3.3 に示す．この旅行ブログエントリでは，旅行ブログエントリ内に“今回は～東の方を中心に散策しました”と書かれていることから，京都に少なくとも１回は訪れた経験のあるブロガーであると判断出来る．このようなブロガーが記述した旅行ブログエントリからは，リピーターが勧めるスポットや，あまり知られていないマイナーなスポットの情報などが発見できる．この他にも，旅行ブログエントリの内容からは，お土産情報や購入場所，旅先で立ち寄った場所についての感想なども知ることが出来る．このように旅行ブログエントリは，観光情報を知りたいユーザに対して有用な情報を提供することができる．

4/29京都行ってきました

（写真1）
（写真2）
日帰りで京都に行ってきました。
普通電車でも2時間程度、新幹線だと35分程度と意外に近いものです。

今回は銀閣寺や南禅寺、哲学の道、平安神宮など東の方を中心に散策しました。

普通に建物や道中も楽しかったのですが、今回の(私の)旅の目的は「よーじやカフェ」の抹茶カプチーノと平安神宮そばにあるカフェのスフレ。
もう、建物よりもそっちに気持ちがいっちゃってました。

「よーじや」はあぶらとり紙が有名で、特に祇園本店が有名ですが、哲学の道にも「よーじや」があります。
そこの「よーじや」はカフェが一緒になっているのです。
面白いカプチーノがあるとのことで以前から気になっていたので早速行ってみると、さすがゴールデンウィーク。満席で少し待ち時間。

と、いうことで店舗のほうで買い物して時間を潰しました。

「よーじや」のあぶらとり紙はよくあぶらがとれるので有名です。
「よーじや」のあぶらとり紙といえば、特徴的な手鏡の顔のキャラクターの表紙に茶色いあぶらとり紙が1冊20枚入りで入っているものが普通ですが、今回のように春に行くと「桜」のあぶらとり紙を販売しています。
といってもただ色がピンクになるだけなんですが(^^;)
とりあえず3冊セットを購入してみました。

そうこうして待つこと30分程度、やっと中に入れました。
中は京都らしく完全和室で、個室になっていて、窓からは和風庭園を見ることができます。

さて、ここで注文したのが、「よーじや」のキャラクター?の顔が書かれたカプチーノ。(1枚目写真)
これが気になっていたのです。

見た目の面白さもさることながら、味も濃厚な抹茶の味と甘みで非常においしかったです。

・・・・・・・・・・・・・・・(省 略)・・・・・・・・・・・・・・・

たまにはきままにのんびり美味しい食べ歩き旅行もいいなぁと思いました。

以上、今回の京都はもうホント甘くておいしい旅でした。

(http://blogs.yahoo.co.jp/akahire_bottle/48159896.html)

図 3.3　旅行ブログエントリの例

3.3 旅行ブログエントリの自動抽出

3.3.1 旅行ブログエントリの自動抽出手法

図 3.3 に示す旅行ブログエントリの例のように，旅行ブログエントリには，“旅行”や“散策”など，旅行の際に頻繁に使用される単語が出現する．このような旅行の際によく使用される語を人手で収集し，旅行ブログエントリの抽出のための手掛かり語として使用する．手掛かり語は，417 語を使用した．使用する手掛かり語の例を，表 3.1 に示す．また，旅行ブログエントリには，旅行先で撮影した写真も多数含まれる．本研究では，手掛かり語の有無，写真の有無を，機械学習の素性として使用する．

表 3.1 旅行ブログエントリの抽出に使用した手掛かり語の例

名詞	旅行，旅，ツアー，観光，見物，散策，自由行動，名所，観光スポット，温泉，寺，景色，土産，名物，買い物，目的，有名，人気，お勧め，記念，満喫，疲れ，無事，写真，ガイド，日程，スケジュール，出発，到着，移動，宿泊，ホテル/宿/旅館，日帰り，飛行機/新幹線/電車/車，など
動詞	歩く/歩き/歩い，行く/行って/行き，買う/買って/買い，探す/探し，食べ，着く/着き/着い，乗る/乗って/乗り，見る/見て/見た，など
形容詞	美しい，素晴らしい，楽しむ，美味しい，など
副詞	いよいよ，いざ，せっかく，どうしても，のんびり，はるばる，ぶらぶら，わくわく，がっかり，など
旅行ブログエントリによく使用される表現	といえば，また行きたい，行ってきます，行ってきました，訪れました，帰ってきました，来ています，着きました，楽しかった，行ってみたかった，など

しかし，すべての旅行ブログエントリに，このような手掛かり語は含まれているわけではない．例えば，あるブロガーがノルウェー旅行について複数のブログエントリにわたって日記を書いていた場合，最初のエントリには“私たちはノルウェーに旅行に行った”と書いてあっても，2 ページ目のエントリには“野生の羊にあったんだ！”としか記述されて

いないこともある．この場合，2 ページ目のエントリには旅行に関連した表現が含まれていないため，2 ページ目のエントリを旅行ブログエントリであると判定することは困難である．そこで，それぞれのターゲットとなるエントリについてのみ見るのではなく，前後のエントリにも注目することで，旅行ブログエントリの抽出を行う．

そこで本研究では，旅行ブログエントリの抽出を，系列ラベリング問題として解き，機械学習を用いて解決する手法を提案する．機械学習の手法には，近年自然言語処理の分野において，実験に用いられ高い精度を示している CRF を使用した．CRF に与える素性とタグは以下のとおりである．我々は予備実験の結果から k=4 と定めた．なお，図 3.4 では、説明のため k=2 の場合を例として示している。

- ターゲットとなるエントリより前の k 個のエントリに付与されたタグ
- ターゲットとなるエントリの前に存在する、ターゲットからの距離が k 以内のエントリに存在する手掛かり語の有無
- ターゲットとなるエントリの後に存在する、ターゲットからの距離が k 以内のエントリに存在する手掛かり語の有無

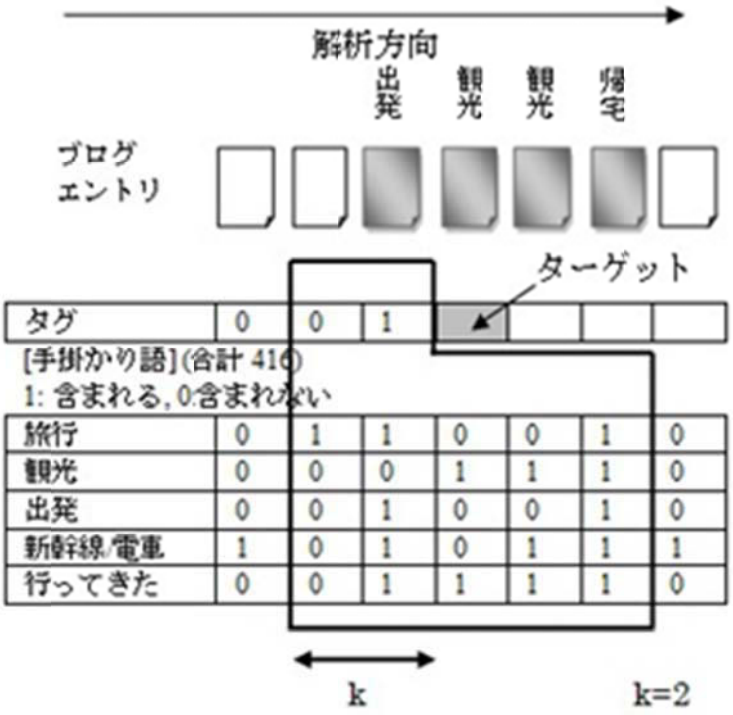


図 3.4　機械学習に与えた素性とタグの例

3.3.2　実験

前章で述べた手法の有効性を評価するため，旅行ブログエントリの抽出実験を行った．

実験に用いるデータ

日本語で書かれた約 1,100,000 エントリから 317 人のブロガーによって書かれた 4,914 エントリをランダムに抽出した．この 4,914 エントリを人手で旅行ブログエントリかどうかを判定した結果を実験に使用した．その結果，“旅行ブログエントリ”と判定されたのは 420 エントリであった．

機械学習と評価尺度

機械学習器には CRF++ *[8]を使用した．また，精度と再現率を用いて評価を行った．4 分割交差検定を行った．

比較手法

旅行ブログエントリの検出を系列ラベリング問題として解く手法を提案した．この提案手法の有効性を確かめるため，比較手法として，前後のエントリの素性を使用せず，注目しているブログエントリのみの素性を使用した旅行ブログエントリの検出を行った．

実験結果と考察

実験結果を表 3.2 に示す．表 3.2 より，提案手法は，精度は 26.2%上がったが，再現率は 13.3%下がった．旅行ブログエントリの抽出の精度が低いと，観光情報を抽出する際の精度が低くなってしまうため，本研究では再現率よりも精度を重要視した．

表 3.2　旅行ブログエントリの自動抽出の結果

手法	精度	再現率
提案手法	86.7	38.1
比較手法	60.5	51.1

人手では“旅行ブログエントリである”と判定したが，提案手法では“旅行ブログエントリではない”と誤って判定したエントリが 266 件存在した．このエントリの中から 50 エ

*[8] http://www.chasen.org/~taku/software/CRF++/

ントリを任意に選び，抽出誤りについて分析を行った．以下に抽出誤りの主要な原因を示す．

(1) 複数エントリにわたる旅行記の一部(50%)
(2) 記載内容が 3 行以下のエントリ(10%)
(3) その他(40%)

以下に，それぞれの検出誤りについて説明する．

(1) 複数エントリにわたる旅行記の一部(50%)

50 件のうち 25 件(50%)が，複数エントリにわたる旅行記の一部であった．複数エントリにわたる旅行記の場合，最初のエントリが“旅行ブログエントリである”と判定できなければ，残りのエントリも“旅行ブログエントリである”と判定することはできない．この抽出誤りの原因は，手掛かり語の不足であった．提案手法では，人手で選択した手掛かり語を使用しているが，手掛かり語を増やす一つの手法として，N グラムを自動的に検出した旅行ブログエントリにあてはめ，手掛かり語を網羅的に集めることで問題を解決することができると考えられる．

(2) 記載内容が 3 行以下のエントリ(10%)

50 件のうち 5 件(10%)が，記載内容が 3 行以下のエントリであった．この抽出誤りの原因は，提案手法で判定するためには短すぎるからだと考えられる．

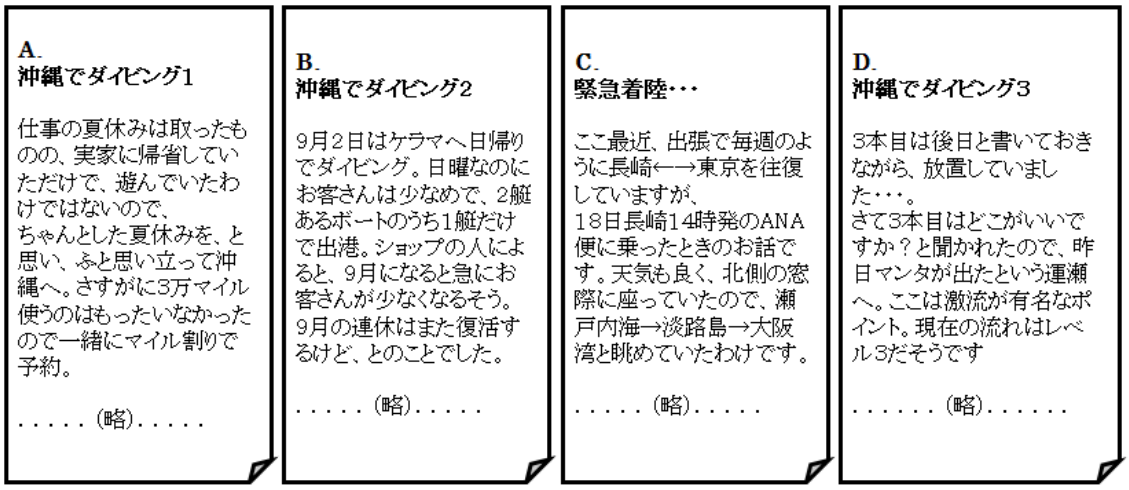
また，人手では“旅行ブログエントリではない”と判定したが，提案手法では“旅行ブログエントリである”と判定したエントリが 26 件存在した．この 26 件の抽出誤りは，大きく次の 4 種類に分類することができる．

(1) エントリの前後に旅行ブログエントリが存在(38.5%)
(2) 地元紹介のエントリ(34.7%)
(3) 他人の旅行を紹介しているエントリ(11.6%)
(4) その他(15.2%)

以下に，それぞれの抽出誤りについて説明する．

(1) エントリの前後に旅行ブログエントリが存在(38.5%)

26 件のうち 10 件(38.5%)が，エントリの前後に旅行ブログエントリが存在していた．エントリの抽出失敗例を図 3.5 に示す．図 3.5 は，あるブロガーの複数エントリにわたる旅行記を示している．この例では，実際には A・B・D のエントリが“旅行ブログエントリである”と判定されなければならないが，ユーザが旅行記を記述する途中で，旅行記の内容とは全く関係のない別のエントリ C を記述してしまっているため，システムでは，誤って C のエントリも旅行記の一部と判断してしまっている．この問題を解決する方法として，エントリのタイトルを前後のエントリと関連づけて判定する（タイトルに“その 1、その 2、1 日目、2 日目”等が含まれる）など考える必要がある．



(http://blogs.yahoo.co.jp/aa37061/51605848.html)

図 3.5　旅行ブログエントリの抽出失敗例

(2) 地元紹介のエントリ(34.7%)

26 件のうち 9 件(34.7%)が，地元住民による地元の紹介エントリであった．この抽出誤りの原因は，ブロガーの居住区情報が反映されていないため，旅行で訪れた場所か，日常生活圏で訪れた場所なのか判定できないためである．これはブログ著者の性別，年齢，居住区などの属性を文体から自動推定する研究の成果を利用することで，解決できるのではないかと考えられる．

(3) 他人の旅行を紹介しているエントリ(11.6%)

26 件のうち 3 件(11.6%)が，他人の旅行を紹介しているエントリであった．自らが体験した旅行についての記事ではないため，人手では“旅行ブログエントリでない”と判定される．しかし，他人の旅行について記事を書いているため，旅行に関する単語が頻繁に出現し，提案手法では“旅行ブログエントリである”と判定されてしまった．

3.4 旅行ブログエントリからの観光情報の自動抽出および有用性評価

旅行ブログエントリは，観光情報の抽出のための有用な情報源であることを確かめるため，旅行ブログエントリから観光情報を自動で抽出する．更に，提案した観光情報抽出手法を，旅行ブログエントリ，一般のブログエントリ，Web 文書に適用して観光情報を抽出し，結果を比較することで旅行ブログエントリの有用性評価を行う．

3.4.1 旅行ブログエントリからの観光情報の自動抽出手法

観光情報としては，土産物，観光名所，宿泊施設，飲食店に関する情報など様々な情報が挙げられるが，本研究では，土産物，観光名所に関する情報を観光情報として抽出する．本節では，旅行ブログエントリから土産物情報と観光名所情報を自動抽出する手法について説明を行う．ここで，土産物情報は，地域名と土産物の対，観光名所情報とは，地域名と観光名所の対であるとする．

土産物情報と観光名所情報の自動抽出には，表層パターンと機械学習を用いる．表層パターンを用いた手法とは，あらかじめ決められたパターンを埋める形でテキスト中の情報を抽出する手法である．そのため，高い精度で情報を抽出できるが，網羅性の点で十分でないという問題がある．そこで，表層パターン手法で抽出されたデータから機械学習用データを自動的に作成し，機械学習を用いることにより，低コストで網羅性の高い観光情報を抽出する．土産物情報の抽出の流れを図 3.6 に示す．

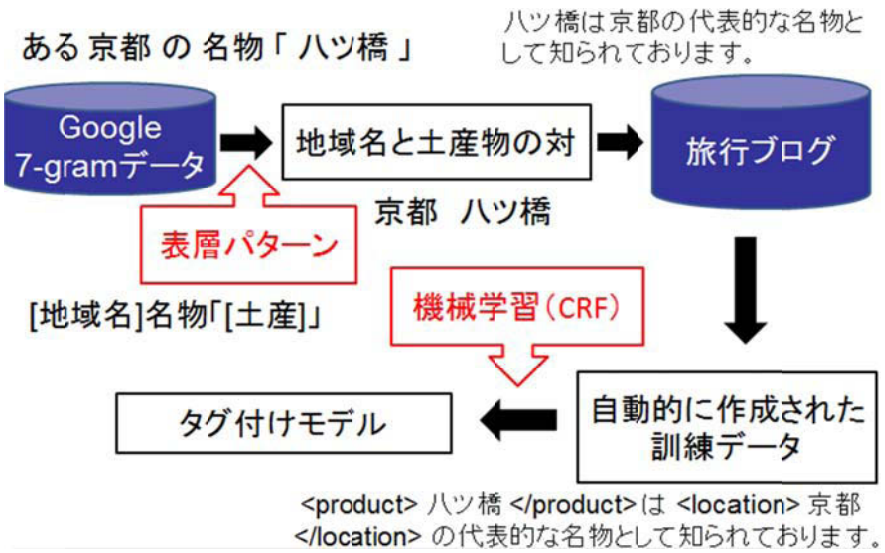


図 3.6　土産物情報抽出の流れ

まず，土産物リスト（地域名と土産物の対），観光名所リスト（地域名と観光名所の対）を作成する．これらのリストは Google から提供されている“Web 日本語 N グラム”データベース に，表層パターンをあてはめ，自動で抽出した．このデータベースは，Web 上にある日本語で書かれた 20 億文から抽出された N グラム(N=1～7)で構成されている．土産物リストの作成には，“[地域名]　名物　「[土産物]」”，観光名所リストの作成には，“[地域名]にある観光名所「[観光名所]」”という表層パターンを使用した．その結果，土産物リストには地域名と土産物の対が 482 対，観光名所リストには地域名と観光名所の対が 35,827 対登録された．

次に，機械学習を使用することで，新しい地域名と土産物の対，地域名と観光名所の対を得る．土産物情報の抽出のための機械学習の訓練用データは，以下の方法で準備する．

Step 1　地域名と土産物両方を含む 200 文を選ぶ．ここで自動的に“location”（地域名）と“product”（土産物）タグを付与したタグ付きの 200 文を生成する．

Step 2　地域名だけを含む 200 文を準備する．また，これらの文に“location”タグを付与したタグ付きの 200 文を生成する．

Step 3　タグ付きの 400 文を機械学習に与え，これらの文に自動的に“location”と“product”タグを付与する．

タグを付与した例を，以下に示す.

<location>広島</location>名物<product>もみじ饅頭</product>をどうぞ。
<location>福岡</location>では銘菓<product>ひよ子</product >。

同様にして，“location”（地域名）と“spot”（観光名所）タグを付与することで，観光名所情報の抽出のための訓練用データを作成した．タグを付与した例を以下に示す.

この週末、<location>新潟県</location><spot>瀬波温泉</spot>へ行ってきました。
ＰＴＡのお友達４人で<location>京都</location>の<spot>嵐山</spot>に行きました。

本研究では機械学習として CRF を使用した。CRF 基本手法は与えられた文に含まれる語を分類するのに使用した。素性とタグは以下のように CRF に与える。

- ターゲットとなる単語から、CRF に与える前後の単語数 k
- ターゲットとなる単語の前に存在する、ターゲットからの距離が k 以内に現れる単語
- ターゲットとなる単語の後に存在する、ターゲットからの距離が k 以内に現れる単語

我々は予備実験の結果から，土産物情報の抽出のとき k=2，観光名所情報の抽出のとき k=4 と定めた．また，機械学習には以下の素性を使用した.

土産物情報の抽出の際に使用した素性

- 単語
- 品詞
- 単語に引用記号がついているかどうか
- 単語が“名物”，“名産”，“特産”，“銘菓”，“土産”のような手掛かり語であるかどうか
- 単語が表層格かどうか
- “ケーキ”や“ラーメン”のような土産物や名産物の名前によく使われる単語が含まれているかどうか

観光名所情報の抽出の際に使用した素性

- 単語
- 品詞

- 単語に引用記号がついているかどうか
- 単語が動詞かどうか
- 単語が表層格かどうか
- “博物館”や“ランド”のような観光名所の名前によく使われる単語が含まれているかどうか
- “市”や“出張”などの不用語が含まれているかどうか

3.4.2 実験および有用性評価

実験に用いるデータ

旅行ブログエントリは，観光情報の抽出のための有用な情報源であることを確かめるため，以下の 3 つの情報源を用いて観光情報を抽出する．それぞれの情報源から，土産物情報・観光名所情報を抽出し，出現頻度によりランク付けを行った。

- 旅行ブログ(提案手法)：3.3 節の手法により，日本語で書かれた約 1,100,000 エントリから，旅行ブログエントリとして検出した 17,266 旅行ブログエントリ中の全ての文(80,000 文)
- 一般ブログ:約 1,100,000 ブログエントリから選択した任意の 80,000 文
- Web 文書:ウェブ 5 億文データベース[27]から選択した任意の 80,000 文

評価尺度

評価尺度としては，上位にランク付けられた土産物情報，観光名所情報に対して，精度を求めた．5 間隔で上位 5 位から 100 位まで精度を計算した．

実験結果と考察

土産物情報・観光名所情報の上位 100 種類の抽出結果を図 3.7, 図 3.8 にそれぞれ示す．土産物情報の抽出において，旅行ブログ手法（提案手法）では 74.0%，Web 文書手法では 7.0%，一般ブログ手法では 20.0%の精度を得た．また，観光名所の抽出において，旅行ブログ手法（提案手法）では 71.0%，Web 文書手法では 37.0%，一般ブログ手法では 31.0%の精度を得た．土産物情報・観光名所情報の抽出において，旅行ブログ手法（提案手法）は，Web 文書手法や一般ブログ手法に比べ，高い精度を得ることができた．よって旅行ブログエントリは，観光情報の抽出のための有益な情報源であるといえる．

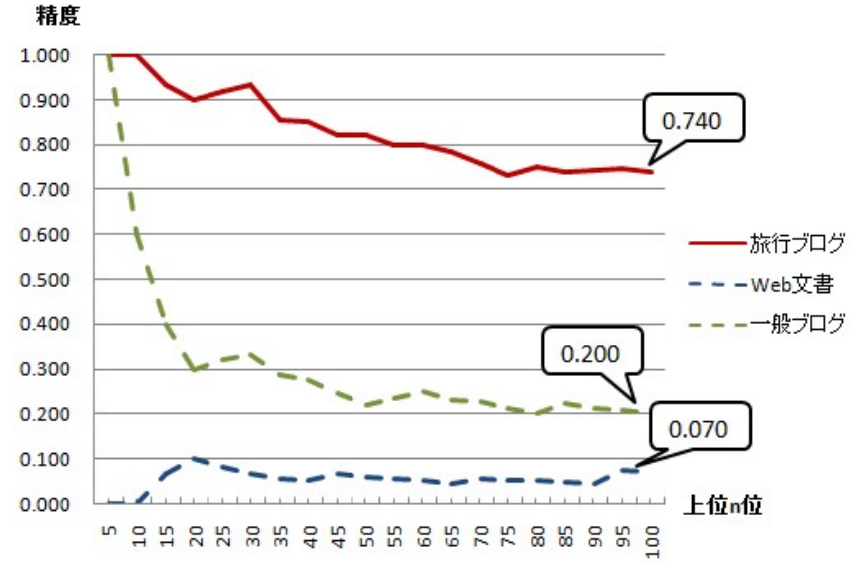


図 3.7　上位 n 位の土産物情報の抽出精度

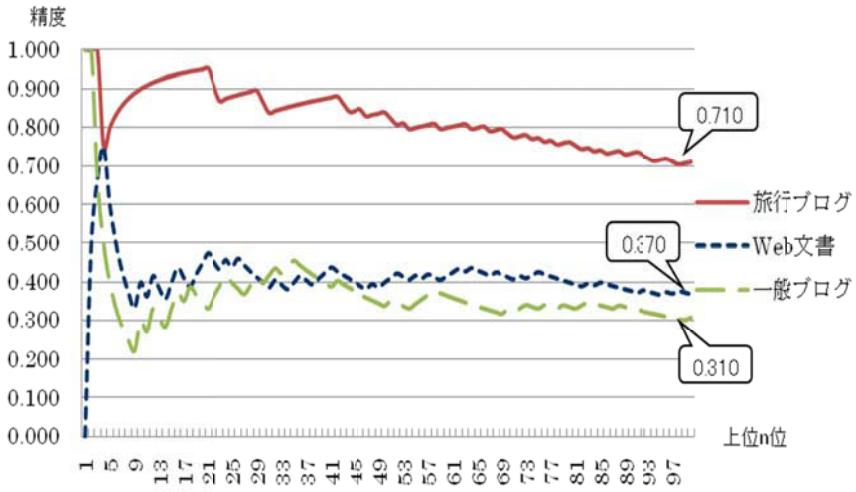


図 3.8　上位 n 位の観光情報の抽出精度

Google N-gram データベースから作成した土産物，観光名所のリストに含まれていないが，本研究で行った各手法により新しく抽出された土産物，観光名所の種類を表 3.3 に示す．

土産物情報の抽出において，旅行ブログ手法では 41 種類の土産物を抽出することができた．一方で，Web 文書手法では 7 種類，一般ブログ手法では 15 種類であった．また，観光

情報の抽出において，旅行ブログ手法では 32 種類の観光名所を抽出することができた．一方で，Web 文書手法では 24 種類，一般ブログ手法では 16 種類であった．これらの結果より，観光情報の情報源として旅行ブログエントリの有益性を示せたといえる．

表 3.3　各手法で新しく抽出された土産物と観光名所の件数

	土産物	観光名所
旅行ブログ(提案手法)	41	32
Web 文書	7	24
一般ブログ	15	16

次に，(1)土産物情報の抽出，(2)観光情報の抽出における抽出誤りについて考察を行う．

(1)　土産物情報の抽出誤り

土産物情報の抽出における提案手法の上位 100 位間の典型的な抽出誤りは，土産物として土産物の販売店の名前を間違えて抽出したことである．これらの店の多くでは土産物を売っている．また，土産物の販売店と土産物は，地域名と土産物と似たパターンで記述されることがあるため，誤って抽出されたと考えられる．この問題は，土産物とその土産物の販売店の対を抽出することで解決できると考えられる．

(2)　観光名所情報の抽出誤り

観光名所情報の抽出における提案手法の上位 100 位間の典型的な抽出誤りについて考察を行う．抽出誤りの例として，イタリアの観光名所である“ピサの斜塔”のように，観光名所に地域名が含まれている場合に，地域名として“ピサ”，観光名所として“斜塔”が抽出されるという誤りがあった．また，地域名“日本”，観光名所“草津温泉”という対のように，誤りではないが地域名の範囲が適切でないものもあった．これは，旅行記を記述したブロガーの居住地域により，地域名の表現が異なるために起こったと考えられる．これは観光情報の利用者によって，地域名の範囲を限定することで解決できると考えられる．

3.5　まとめ

本章では，ブログデータベースから旅行ブログエントリを自動抽出するための手法を提案した．旅行ブログエントリの自動抽出では，精度 86.7%，再現率 38.1%を得た．

また，旅行ブログエントリから，観光情報として土産物情報と観光名所情報を抽出する手法を述べた．抽出された上位 100 件の土産物において精度 74.0%，抽出された上位 100 件の観光名所において精度 71.0%を得ることができた．また，旅行ブログエントリの観光情報の情報源としての有効性を確認するため，一般ブログ，Web 文書からの抽出した結果との比較を行った．実験の結果，旅行ブログエントリの観光情報の情報源としての有用性を確認した．

また，これらの手法と，ブロガーの属性（性別、居住地域など）を抽出する技術を利用することで，利用者に適した観光情報を提供することができようになると考えられる．

第4章　旅行ブログエントリを使用した観光情報リンク集の自動構築

Venice, Italy, 2012

本章では，旅行ブログエントリを利用した，観光情報リンク集を自動構築する手法について述べる．4.1 節では関連研究，4.2 節では観光情報リンクの検索システムの動作例を紹介する．4.3 節では観光情報リンクの自動収集手法について説明し，4.4 節で実験と考察を行う．4.5 節でまとめる．

4.1　リンクの自動分類に関する研究

旅行ブログエントリ中に含まれるリンクを分類することで，観光情報リンク集を自動構築する．本節では，「リンクの自動分類」に関連する研究を紹介する．

本研究では，旅行ブログエントリ中に含まれるリンクを抽出し，「食べる」や「見る」などの観光に特化したタイプに分類することで，観光情報リンクを自動収集する．ブログ中に含まれるリンクを分類する研究としては，Kale ら[28]，Ishino ら[29]，Martineau ら[30]の研究がある．

Ishino らや Kale らは，リンク周辺文字列に含まれる評価表現の比率により，リンクの極性判定を行っている．リンクの評価極性とは，このリンク先ブログに対するリンク元ブログの評価を表すものである．Ishino らの，リンクの極性判定の手順を以下に手法に示す．

1. ブログ中のリンク先ブログに対する評価が書かれている文（引用個所）を抽出する．
2. 引用個所に含まれる評価表現の比率により，リンクの評価極性を判定する．

引用個所の抽出は，手掛かり語を用いたルールにより抽出している．本研究では，Ishino らと同様に，手掛かり語を用いたルールにより引用個所の抽出を行い，その引用個所を用いてリンクの分類を行う．Ishino らは，リンクの評価極性をポジティブとネガティブに分類しているが，本研究では，リンクは 4.3.3 節で説明する旅行に特化したタイプに機械学習を用いて分類する．

Martineau らは，機械学習を用いていることで，ブログ中のリンクについて，以下の 3 つの観点から分類を試みている．Martineau らは提案する次元を多分類タスクとしてテストしており，実験では単語 uni-gram を素性とし，SVM を用いて分類器を構築している．

- なぜ，著者はリンクを張るのか？
- 著者は，正確には何を指摘しているのか？
- そのことについて，著者はどのように感じているのか？

本研究では，Martineau らの手法と同様に，機械学習を用いてリンクの分類を行う．本研究では，観光情報に特化したリンク集の構築を目標としているため，リンクは 3.3.3 節で説明するタイプに分類する．

本研究では，旅行ブログエントリから収集したリンクに対し，タイプ判定を行うことで，

観光情報リンク集の構築を行った．観光情報リンク集を利用者に提示する場合に，スニペットをどのように生成するかという問題点がある．本研究では，リンクの分類を行う際に利用した引用個所をスニペットとして表示するようにした．このスニペットを読むことで，リンク先サイトに関する感想などの情報も得ることができるようになっている．

文書間のリンクを，手掛かり語に基づいて分類する研究に，難波の研究[31]がある．学術論文中には，当該論文と被参照論文との関係について記述されている個所(引用個所)がある．難波は，この引用個所を抽出し，抽出された引用個所から参照タイプを決定する．

引用個所の抽出とは，参照の出現する段落において，参照のある文と文間のつながりが強いと考えられる文を参照の前後の文から抽出処理であると考え，「引用個所候補となる文の前後の文に手がかり語が出現すれば，その文も引用個所候補に含める」というルールを用いて引用個所の自動抽出を行っている．このようにして抽出された引用個所から得られる情報を，難波の研究では参照情報と呼んでいる．難波の研究はこの論文間の参照情報に着目している．引用個所からは，被参照論文の重要点や当該論文と被参照論文との相違点を明示する有用な情報が得られる．また，引用個所を読めば参照の理由(参照タイプ)が分かる．参照タイプは次のように分類できる．

- type C (問題点指摘型)
  新しく提案した理論や，構築したシステムの新規性について述べる場合，関連研究との比較，あるいは既存研究の問題点の指摘を行う場合がある．このような目的の参照タイプを type C と呼ぶ.
- type B (論説根拠型)
  新しい理論を提唱したり，システムを構築する場合，他の研究者の研究の成果を利用する場合がある．このような参照タイプを type B と呼ぶ.
- type O (その他型)
  Type B にも type C にもあてはまらない参照を type O（その他型）と呼ぶ.

こうした参照情報から，当該論文の関連論文の中での位置づけが明らかになるため，特定分野の研究動向の概要の把握に有用である．難波の研究では，引用個所を抽出することで，当該論文の作者がどのような理由で被参照論文を参照したかを明らかにしている．本研究でも，難波と同様に引用個所を抽出することにより，リンク元ブログの作者が，リンク先サイトをどのような目的で利用したのかを判定する．

4.2 観光情報リンクの検索システムの動作例

観光情報リンクの検索システムについて，その動作例を紹介する．図 4.1 は，観光情報リンクの検索システムの画面である．以下では，あるキーワードに関する観光情報リンクを検索する場合の一般的な操作手順にについて説明する．まず，画面上部の検索窓に（図中①）に、キーワードを入力する。（図 4.1 の場合「大阪」というキーワードが入力されている）．この状態で，「search」ボタン（図中②）を押すと，キーワードに関連する観光情報リンクが表示される．観光情報リンクは，自動で収集した旅行ブログエントリから抽出されたリンクである．旅行ブログエントリ中に記載されている観光情報リンクには，観光名所に関するリンク，ホテルに関するリンク，食事に関するリンクなど様々なリンクが存在する．そのため，本研究ではリンク先サイトの種別を自動で分類する．本研究では，旅行ブログエントリから観光情報リンクを抽出し分類する手法を提案し，実験によりその有効性を検証する．

図 4.1　観光情報リンク検索システムの動作例

4.3 観光情報リンクの自動収集

本節では，旅行ブログエントリから観光情報リンクを自動収集する手法について説明を行う．旅行ブログエントリには図 4.2 に示す旅行ブログエントリのようにリンクを含んでいるものがある．図 4.2 中の「http://www.yudaonsen.com/index.html」が，図 4.3 の湯田温泉の Web サイトへのリンクである．このように，旅行ブログエントリ中には旅行の際に有益なサイトへのリンクが多く含まれると考えられる．そこで本研究では，旅行ブログエントリ中に存在するリンクを収集し，4.3.3 節で示すリンクタイプに分類することで，観光情報に特化した観光情報リンク集の構築を行う．

てなわけで、九州を脱出して通勤割引最長区間の山口県は美祢市に到着。
秋吉台から、目的地の湯田温泉に向かう夜の山道はホタルが飛んでたかも？
このあたりは源氏ボタルの有名な生息地なんだとか...

また、今年もそんな季節になったよなぁ。。。
そうだ！　久しぶりに「ホタル見学ツアー」に行ってみよう！

しかし今回も、とってもいい温泉だった。
山口県「湯田温泉」
http://www.yudaonsen.com/index.html

図 4.2　リンクを含む旅行ブログエントリの例

図 4.3　図 4.2 の旅行ブログエントリのリンク先のサイト

4.3.1　観光情報リンク集の自動構築の手順

観光情報リンク集構築の手順を以下に示す．本研究では，‘。’や‘.’があれば改行を入れるという簡単なルールにより文分割を行っている．Step 2 の引用箇所の抽出は 4.3.2 節，Step 3 のリンクタイプの判定は 4.3.3 節，4.3.4 節で述べる．

Step 1　旅行ブログエントリのテキストを入力する．
Step 2　入力テキストから観光情報リンク部分を見つけて，その観光情報リンクに関する情報が記述されている文（引用個所）を抽出する．
Step 3　引用個所を用いてリンクタイプの判定を行う．

4.3.2　引用箇所の抽出

引用個所の抽出について説明を行う．観光情報リンクに関する情報は，観光情報リンクの周辺に記述される傾向があるが，観光情報リンクから離れた場所にも記述される場合もある．よって本研究では，手掛かり語により，引用個所を自動で抽出する．サイトを紹介する際には，リンク先サイトのタイトルが“「」”や“『』”などの記号で囲まれている場合がある．また，“紹介”，“の HP”などの語が使われるため，これらを手掛かり語として使用した．以下に，手掛かり語と，手掛かり語を用いた引用個所の抽出ルールを示す．

手掛かり語　（26 個）

- “「」”，“『』”などリンク先サイトのタイトル周辺に使用される記号　（6 個）
- “紹介”，“の HP”，“公式サイト”，“こちら”などリンク先サイトを紹介する際に使用される単語　（20 個）

引用個所の抽出ルール

1. 観光情報リンクが含まれている文を抽出する．
2. 観光情報リンクが含まれている文の前後 X 文を抽出する．（予備実験より X=2 とする．）
3. 観光情報リンクが含まれている文，その観光情報リンク前後 X 文に，手掛かり語（記号または単語）が含まれていれば，手掛かり語の周辺文字列を，リンク先サイトを指し示す語（Keyword）として抽出する．例えば、“『A 公園』”，“B 公園の HP”という文があれば，Keyword はそれぞれ“A 公園”，“B 公園”となる．
4. Keyword が含まれている文を抽出する．

図 4.4 の旅行ブログエントリを用いて，引用個所抽出ルールを説明する．ルール 1 により，観光情報リンクが記述されている 7 文目を引用個所として抽出する．次に，ルール 2 により，観光情報リンクが含まれている文の前後 2 文（5，6，8，9 文目）を引用個所として抽出する．ルール 3 により，6 文目に“の HP”という手掛かり語が含まれているため，“の HP”の直前の単語である“バガテル公園”を Keyword とする．ルール 4 により，Keyword である“バガテル公園”という単語が含まれている 2，10 文目を引用個所として抽出する．よって，図 4.4 の旅行ブログエントリから抽出される引用個所は，2，5，6，7，8，9，10 文目である．なお，機械学習による引用個所の抽出も試みたが，よい結果が得られなかったため，本手法ではルールにより引用個所を抽出することにした．

1　チェックアウト後、いつものようにパパ&ママの寄り道が始まります!!
2　ということで、まずは河津の【バガテル公園】に行ってきました☆
3　四季の蔵から、車で数分圏内にあります。
4　ワンコもお散歩 OK なので、犬連れには嬉しい場所です
5　メッチャ、綺麗でしたよ～♪
6　※バガテル公園の HP は、こちら→
7　<ahref=”http://www.bagatelle.co.jp/index.html” target=_blank>
　　http://www.bagatelle.co.jp/index.html
8　↑いうまでもなく、美しいバラの数々(写真)
9　四季の蔵の朝ごはんがボリューム満点だから、これくらいで充分です!!
10　初めて来たバガテル公園ですが、ワンコ OK だし、
11　季節によってはお花が綺麗なのでいいかも～♪
12　ランチメニューも充実しているし、また今度も来ようっと(ノ▽≦*)キャハッッッ♪

図 4.4　リンクを含む旅行ブログエントリの例

4.3.3　リンクタイプの定義

リンクタイプは以下のように判定する。

- S (Spot) ：　旅行者が訪れた名所，施設に関する情報(歴史，生息する動物など)かどうか.
- H (Hotel) ： 旅行者が宿泊したホテルや宿に関する情報かどうか.
- R (Restaurant) :　旅行者が食事をとったレストラン，食べ物，食べ物を販売するお店に関する情報かどうか.
- O (Other) :　S，H，R のいずれにも判定されないもの.

次に，人手により各リンクタイプに判定されたリンクの例を示す．図 4.5 に示す旅行ブログエントリは，広島県にある備北丘陵公園について記述された旅行ブログエントリである．リンク周辺の文字列から，この旅行ブログエントリに含まれているリンクは備北丘陵公園を紹介するサイトへのリンクであることがわかるため，観光名所に関する情報を紹介するサイトへのリンクであるといる．よって，このリンクは人手によりリンクタイプ S であると判定される.

国営備北丘陵公園（びほくきゅうりょうこうえん）のイルミネーション
を見に行って来ました！
http://www.bihoku-park.go.jp/
備北丘陵公園と言えば、以前にドリカムやミスチルやモー娘や
スマップが野外コンサートをしたことがある場所＾＾
公園の広さは３６０ｈａ。。。とにかく広い！
中国自動車道の庄原インターから車で５分の位置にあります♪

図 4.5　人手によりリンクタイプが S であると判定されたリンクの例

図 4.6 に示す旅行ブログエントリは，ディズニーランドホテルについて記述された旅行ブログエントリである．リンク周辺の文字列から，この旅行ブログエントリに含まれているリンクは，東京ディズニーランドホテルを紹介するサイトへのリンクであることがわかる．よって，人手によりリンクタイプ H であると判定される．

宿泊した「東京ディズニーランドホテル」のことをお話します～♪
ホテルは天井が高く素敵なロビー！
私たちの部屋は
ピーターパンルーム
キャラクターのお部屋というものの
おちついたインテリアが素敵です。

図 4.6　人手によりリンクタイプが H であると判定されたリンクの例

図 4.7 に示す旅行ブログエントリは，ドルチェというジェラード店を訪れた際の旅行記である．リンク周辺の文字列から，ジェラード店「ドルチェ」というお店を紹介するサイトへのリンクだということがわかる．よってこのリンクは，食に関する情報を紹介するサイトへのリンクであるとわかるため，人手によりリンクタイプ R であると判定される．

で、３時のデザートは某百貨店地下にあるジェラード店「ドルチェ」でジェラードをいただきました。
「ドルチェ」は広島県生口島瀬戸田にあるお店です
http://www.setoda-dolce.com/
昨日はイタリアンバニラ。本日はチョコチップを食べました。
明日は杏仁か伯方島の塩をいただこうかしらん？？？
と、本日は食べ物ばかりの記事でした。
明日で長かった広島出張も終わりやでぇ～♪

図 4.7　人手によりリンクタイプが R であると判定されたリンクの例

図 4.8 に示す旅行ブログエントリは，餃子スタジアムを訪れた際の旅行記である．餃子スタジアムは，食を売りにした観光スポットであるため，リンクタイプは S と R 両方に判定される．このように各リンクは，複数のタイプに判定される場合もある．

今月はじめに、大阪の学会にいったついでに、浪花餃子スタジアムに行ってきました。
ここは梅田にある、ナムコが運営する餃子テーマパーク建物の３F までのぼっていくと
9 店舗の餃子屋さんが待っています。
私が今回立ち寄ったのは・・・・
久留米屋台餃子・満州が一番
なんでも某 TV 番組で餃子日本一に選ばれた店だとか
. . . . . （略）. . . . . .
餃子スタジアム
 http://www.namco.co.jp/tp/naniwagyoza/

図 4.8　人手によりリンクタイプが S と R であると判定されたリンクの例

リンクタイプが，S，H，R のいずれにも分類されないものを O と判定する．O と判定されたリンクの例には，図 4.9 に含まれている車を運転する際のモラルを掲載したサイトへのリンクなどがある．

高速使わず延々と中国山脈を南下。
標高の高い広島県の山間部は夜風も寒いっ！
. . . . . （略）. . . . . .
高速道路の追い越し車線で、トラックが前の車に追いついたとき出す右ウインカー
前の車をあおっている？と考えるか
排気ブレーキの減速してますの後続車への合図とか。
パッシングに関しては、考えただけでもいっぱいある。
↓こちらのサイトは、解りやすく解説されてておもしろい。
「間違った認識」
<a href="http://homepage1.nifty.com/takapapa/car_5.htm"
TARGET=_blank>http://homepage1.nifty.com/takapapa/car_5.htm#ゴーンの誤解

図 4.9　人手によりリンクタイプが O であると判定されたリンクの例

4.3.4　リンクタイプの判定

本研究では，機械学習によりリンクタイプの判定を行う．学習には，「引用個所に出現する各単語」，「手掛かり語の有無」を素性として与える．

リンクタイプ S のリンク周辺には，観光名所の名前や，“観光”，“見学”，“訪れる” など，旅行者が観光名所に訪れた際によく使う単語が頻繁に出現すると考えられる．このような単語をリンクタイプ S の手掛かり語として収集しリストを作成した．観光名の名前は，Wikipedia などの Web ページから収集した．R，H についても同様の観点から手掛かり語の収集を行った．S，H，R の手掛かり語を，表 4.1，表 4.2，表 4.3 に示す．

表 4.1　リンクタイプ S の手掛かり語

手掛かり語	単語数
Wikipedia から収集した観光名所の名前 (例)　東京ディズニーランド，ユニバーサル・スタジオ・ジャパン， 原爆ドーム，厳島神社，ロマンチック街道，ナイアガラの滝　など	17,812
観光名所の名前に使用される単語 (例)　動物園，博物館，美術館，水族館，遺跡，公園，牧場，寺，城， タワー，リゾート，ビーチ，ランド，パーク，ワールド　など	138
観光の際に使用される単語 (例)　観光，観戦，見学，見物，散策，散歩，移動，参加，参拝， 訪れた，訪問する，行ってきた，立ち寄る，向う，出かける　など	172
その他 (例)　名所，観光地，目的地，写真，撮影，景色，夜景，夕焼け　など	131
合計	17,812

表 4.2　リンクタイプ H の手掛かり語

手掛かり語	単語数
宿泊施設の名前に使用される単語 (例)　旅館，宿，荘，休暇村，リゾート，ホテル，ペンション， ロッジ，コテージ	9
宿泊施設の構成要素 (例)　客室，和室，洋室，和洋室，部屋，ダブル，ツイン，シングル， 温泉，大浴場，フロント，フロアー，ロビー，ヴィラ　など	29
宿泊する際に使用される単語 (例)　泊る，連泊，宿泊，チェックイン，チェックアウト，一泊， 予約，滞在，一晩　など	14
その他 (例)　快適，ゆっくり，まったり，空室，空き　など	21
合計	73

表 4.3　リンクタイプ R の手掛かり語

手掛かり語	単語数
Wikipedia から収集した料理名 (例)　焼肉，牛丼，刺身，餃子，卵焼き，冷麺，炒飯，白玉団子， ローストビーフ，ハンバーグ，オムレツ，ケーキ　など	2,779
Wikipedia から収集した料理の種類 (例)　郷土料理，沖縄料理，京料理，懐石，鍋料理，魚料理，広東料理， ファーストフード，フランス料理，イタリアン　など	114
食事をとる施設の名前に使用される単語 (例)　食堂，亭，喫茶，居酒屋，飲み屋，甘味処，茶屋，横丁， レストラン，リストランテ，カフェ，ダイニング，バー　など	21
食事をとる際に使用される単語 (例)　食べる，召し上がる，試食，飲む，おいしい，旨い，まずい， 辛い，甘い，お腹いっぱい，満腹　など	52
食べ物を指す単語 (例)　ご飯，料理，定食，食べ放題，おやつ，モーニング，ランチ， ディナー，フルコース，メニュー，バイキング，ビュッフェ　など	31
その他 (例)　パティシエ，レシピ，土産，ミシュラン，トッピング　など	31
合計	3,028

4.4 実験

前章で述べた手法の有効性を評価するため，観光情報リンク集の実験を行った．

実験に用いるデータ

3 章の手法により，日本語で書かれた約 1,100,000 エントリから，旅行ブログエントリとして 17,266 件のエントリを検出した．これらの旅行ブログエントリには，7,421 件のリンクが含まれていた．リンクの中には，Wikipedia やブログ，ニュースサイトへのリンクなど，リンク先 URL からリンク先サイトを判定することができるものも含まれている．よって本

研究では, そのようなリンクを除外した 4,155 件のリンクから, 1,000 件のリンクを抽出し, 人手でリンクタイプの判定を行った結果を機械学習に用いる．人手でリンクタイプの判定を行った結果を，表 4.4 に示す.

表 4.4　1,000 件のリンクに含まれる各リンクタイプの件数

リンクタイプ	S	H	R	O
リンク件数	353	98	343	250

比較手法

提案手法の有効性を確かめるため，リンクが含まれている文の前後 X 文を引用個所として比較実験を行う.

- 提案手法
  引用個所の抽出ルールにより、抽出された文を引用個所として使用。

- 比較手法
  リンクの前後 X 文を、引用個所として使用。

機械学習と評価尺度

リンクタイプの判定の学習には TinySVM を用いた．2 次の多項式カーネルを使用し，4 分割交差検定を行った．また，精度と再現率を用いて評価を行った.

実験結果と考察

実験結果を表 4.5 に示す. 比較手法では引用個所として, リンクの前後 X 文を使用した. ここでは，最も実験結果の良い X=2 のときの結果を示す.

表 4.5　リンクタイプの判定の実験結果

リンクタイプ	提案手法		比較手法	
	精度	再現率	精度	再現率
S	72.7	62.5	64.7	54.5
H	81.3	64.9	79.8	63.3
R	76.7	71.9	76.0	72.3
O	48.6	71.6	42.2	59.2

上記の実験結果より，比較手法に比べ，提案手法のリンクタイプ R の再現率が若干低下したが，その他では，提案手法が精度・再現率ともに，高い数値を記録することができた．特に，リンクタイプ S において，精度 8.0 ポイント，再現率 8.0 ポイントの改善を行うことができた．よって提案手法の有効性を示せたといえる．

提案手法，比較手法ともに，他のリンクタイプに比べ O の精度が低くなってしまったのは，S，H，R のいずれにも判定されないものを O としたためである．S，H，R の更なる精度の向上により，O の精度も改善できると考えられる．

次の各段階に分けて，提案手法を用いた際の，リンクタイプの判定誤りの原因ついて考察を行う．

(1) 引用個所の抽出
(2) リンクタイプの判定

以下に，それぞれの段階について説明する．

(1) 引用個所の抽出

本研究では，まず旅行ブログエントリから引用個所を抽出し，リンクタイプの判定を行っている．そのため，引用個所の抽出に誤りがあった場合に，リンクタイプを正しく判定することができないものがあった．引用個所の抽出には，以下の 2 つの問題がある．

(1-1) 引用個所の抽出不足

本研究では人手で収集した手掛かり語を用いて，引用個所の抽出を行った．しかし，リンクに関する情報が記述されている文を，引用個所として抽出できていない場合があった．

1 つ目の原因として，リンクの紹介方法がブログ著者により大きく異なるため，人手で収集した手掛かり語では対応できなかったことが挙げられる．これは，リンク周辺に出現す

る単語を収集し，手掛かり語を網羅的に集めることで解決できると考えられる．

2 つ目の原因として，手掛かり語に頼った抽出手法では，文間の語彙的なつながりを見つけることが困難であることが挙げられる．これは，引用個所の抽出に，語彙的伝搬の情報を加えることで解決できると考えられる．

(1-2) 引用個所の過抽出

旅行ブログエントリにリンクが連続して出現している場合，他のリンクに関する情報を，ターゲットとしているリンクの引用個所として抽出する場合があった．また，リンクの直前や直後に，リンクに関係のない記述があるときに，その文を引用個所として抽出してしまう場合があった．このような場合は，他のリンクとの距離や，リンクの前後の文の語彙的なつながりを考慮に入れることで解決できると考えられる．

(2) リンクタイプの判定

次に,各リンクタイプにおける判定誤りについて考察を行う.人手では“リンクタイプ X”と判定したが，提案手法では，“リンクタイプ X でない”と誤って判定したリンクについて考察を行う．（X=S，H，R）以下に，判定誤りの主要な原因を示す．

(2-1) リンク先サイトに関する記述の不足

判定誤りの主な原因として，リンク先サイトに関する記述が少ない場合に，判定を誤っていた．本研究では，手掛かり語を用いた手法を提案したが，リンク先サイトに関する記述が不足していると，手掛かりとなる語が含まれておらず，提案手法では正しく判定できなかったと考えられる．

(2-2) 手掛かり語の不足

本研究では，人手により収集した手掛かり語を用いて，リンクタイプの判定手法を提案した．リンクタイプの判定誤りの原因として，手掛かり語の不足が考えられる．例として，リンクタイプを R と判定する場合を挙げる．

リンクタイプを R と判定する際の手掛かり語として，“おいしい”など食事をとる際に使用される単語を使用した．しかし本研究では，旅行ブログエントリを情報源として使用しているため，同じ“おいしい”という意味でも“おいしー”，“おいし～”，“美味しい”，“オイシイ”など様々な記述が存在する．このため，人手により手掛かり語を網羅的に収集するのは困難である．この問題を解決する手法として，レストランの口コミサイトなどの口コミを利用することで，より多くの手掛かり語を収集することが考えられる．

人手では“リンクタイプ X でない”と判定したが，提案手法では，“リンクタイプ X”と誤って判定したリンクについて考察を行う．(X=S，H，R) 以下に，判定誤りの主要な原因を示す。

(2-3) 周辺施設に関する記述が存在

リンク先サイトを紹介する際に，タイプの異なる周辺施設を紹介する記述が存在した場合に，判定を誤っている場合があった．例えば，リンク先サイトがホテルに関するサイトであり，人手でリンクタイプは S でないと判定されていたとする．このとき，リンク周辺に，ホテルの部屋から眺める観光名所に関する情報が記述されていた場合に，リンクタイプ S であると誤って判定されていた．

(2-4) 手掛かり語の重複

本研究では，各リンクタイプのリンク周辺に出現しやすい単語を，人手で収集し手掛かり語として使用した．リンクタイプ S の手掛かり語として，“訪れた”という語を登録している．しかし，“訪れた”という語は，レストランに食事に行った際にもよく使われる単語であるため，リンクタイプ R の手掛かり語としても登録している．そのためリンク周辺に“訪れた”という記述があった場合に，誤って判定されてしまった．この問題は，リンクタイプを判定したリンク周辺に出現する N グラムを使用し，各リンクタイプに特化した手掛かり語を自動的に収集することで解決できると考えられる．

4.5 まとめ

本章では，旅行ブログエントリからリンクを抽出し，リンクタイプを判定することで観光情報リンク集を構築する手法を提案した．また，実験により提案手法の有効性を確認した．

本研究では，まず旅行ブログエントリに含まれる観光情報リンクを抽出し，引用個所に含まれる単語を用いてリンクタイプを判定する手法を提案した．引用個所は手掛かり語を用いて自動抽出を行った．また，リンクタイプの判定は，引用個所に含まれる手掛かり語により判定した．リンクタイプ S において，精度 72.7%，再現率 62.5%，リンクタイプ H において, 精度 81.3%, 再現率 64.9%, リンクタイプ R において, 精度 76.7%, 再現率 71.9% を得ることができ，提案手法の有効性を確認することができた。

それぞれの手法では，観光名所の名前や，料理名などを手掛かり語として使用した．しかし，この手掛かり語リストは主に人手で作成しているため，もれが出てくる可能性がある．そのため，今後ブログを中心とした観光情報の組織化を行うに当たっては，観光名所の名前や料理名や土産物の名前を体系的に収集していく必要があると考えられる．

第5章　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

New York, United States of America, 2013

本章では，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について述べる．5.1 節では研究の目的，5.2 節では情報拡張された旅行ガイドブックの閲覧システムの動作例を紹介する．5.3 節では関連研究，5.4 節では旅行ガイドブックを情報拡張するための手法について説明し，5.5 節で実験と考察を行う．5.6 節でまとめる．

5.1 旅行ガイドブックの情報拡張の目的

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．

5.2 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 5.1 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 5.1 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*9．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 5.2 は，図 5.1 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*10のような宿泊施設の予約サイトや，食べログ*11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

*9 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).
*10 http://travel.rakuten.co.jp/
*11 http://tabelog.com/

図 5.1　情報拡張された旅行ガイドブックのページの例

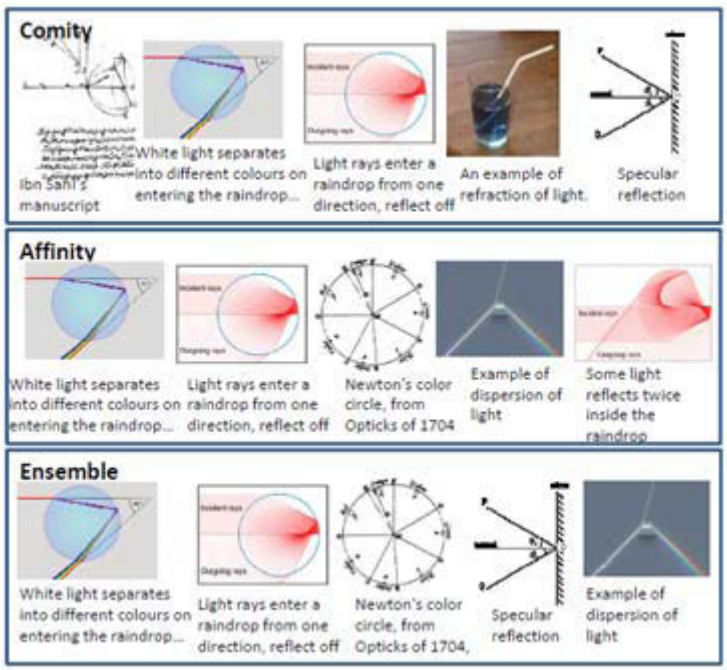
図 5.2　図 5.1 の旅行ガイドブックに自動的に対応付けられた質問応答コンテンツの例

5.3 関連研究

旅行ガイドブックの情報拡張に関連する研究として，「文書の情報拡張」，「タイプ分類」に関連する研究を紹介する．

5.3.1 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[32]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 らの手法により，教科書に対応付けられた画像を図 5.3 に示す．図 5.3 は，科学の教科書の「人間の目と色彩豊かな世界」の章に関連する画像である．この章では，虹の形成と白色光と空の青の分割を含む自然界の光学現象を説明している．



((Rakesh 11)[32]より抜粋)

図 5.3　教科書に対応付けられた画像

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*[12]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*[13]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例を図 5.4 に示す．

http://lab.kn.ndl.go.jp/nii/browse/JPNO45019705/7

図 5.4　電子読書支援システムの例

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

12 http://lab.kn.ndl.go.jp/cms/
13 http://lab.kn.ndl.go.jp/nii/

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[33, 34]. 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．

5.3.2　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[7]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[35]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，5.4.1 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[36]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 [37]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[38]. Yang ら[39]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．

5.4 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 5.4.1 節，5.4.2 節，5.4.3 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．

5.4.1 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 5.1 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 5.1　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 5.5 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 5.5 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．

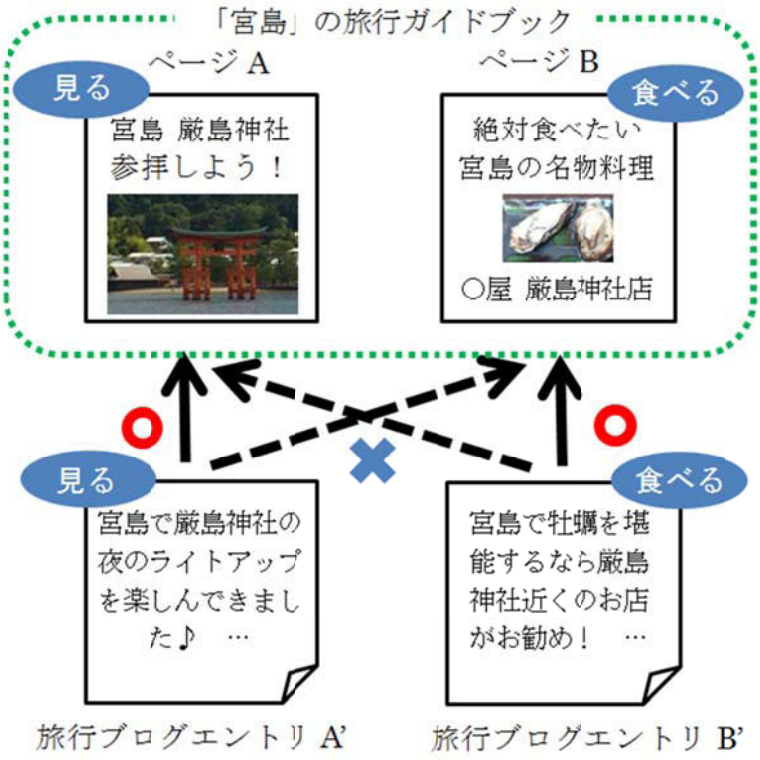


図 5.5　旅行ガイドブックのページへの旅行ブログエントリの対応付けのイメージ

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*[14]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 5.2 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．旅行ブログエントリ，質問応答コンテンツから収集した手掛かり語の例を表 5.3，表 5.4 に示す．

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

表 5.2　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

表 5.3　旅行ブログエントリ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	公園，美術館，花，建築，紅葉，野球
体験する	温泉，露天風呂，アトラクション，大会
買う	買う，土産，買い物，パッケージ，雑貨
食べる	食べる，美味しい，デザート，味，お好み焼き
泊まる	部屋，ホテル，泊まる，浴場，宿泊
その他	徹底的，朝一番，運転，関西空港，趣味

表 5.4　質問応答コンテンツ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	見る，ホテル，花火大会，花火，祭り
体験する	プール，乗り物，アトラクション，海水浴場
買う	買う，売る，土産，お菓子，店
食べる	食べる，おいしい，うまい，ラーメン，食事
泊まる	ホテル，泊まる，宿，旅館，部屋

§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*15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．

5.4.2 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

*15 http://blogs.yahoo.co.jp/

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[40]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 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[41]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 5.1 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 5.6 に示す．図 5.6 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[42]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．

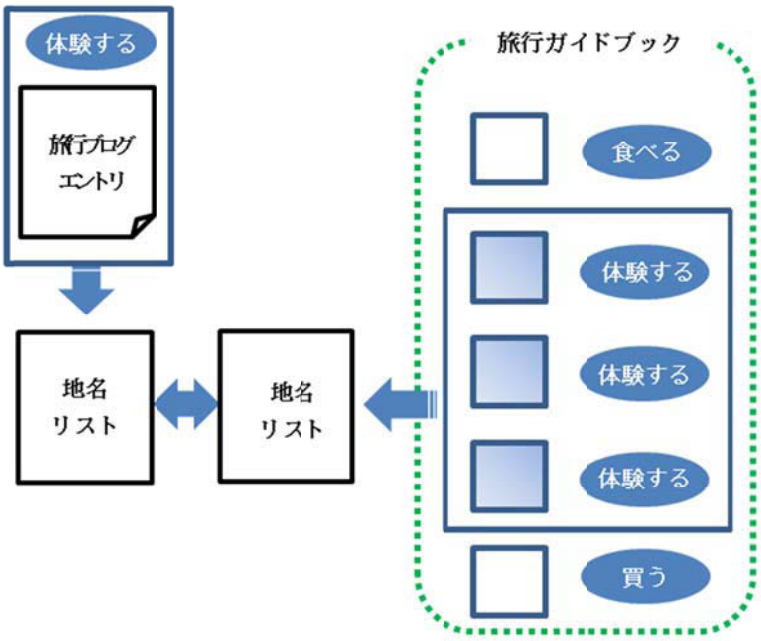


図 5.6　タイプを使用した旅行ガイドブックへの旅行ブログエントリの対応付け

5.4.3　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．

5.5 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，5.5.1 節，5.5.2 節，5.5.5 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 5.5.4 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

5.5.1 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

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 5.5 に示す．

表 5.5　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 5.6，表 5.7，表 5.8 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．

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

表 5.7　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	精度(%)	68.4	55.3	50.7	75.1	49.8	66.4
（比較手法）	再現率(%)	60.6	37.0	20.8	67.4	19.2	48.7
IG	精度(%)	66.7	60.2	54.9	77.2	58.9	65.9
（提案手法）	再現率(%)	64.0	33.7	31.8	69.9	34.3	51.0

表 5.8　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	精度(%)	72.7	56.6	77.5	72.2	82.0	70.9
（比較手法）	再現率(%)	70.6	43.3	41.7	69.6	65.4	61.1
IG	精度(%)	71.1	81.4	90.1	71.8	90.8	80.7
（提案手法）	再現率(%)	44.4	11.9	33.3	47.5	20.4	32.4

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 5.9 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 5.9　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 5.9 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 5.7 に示す. 図 5.7 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.

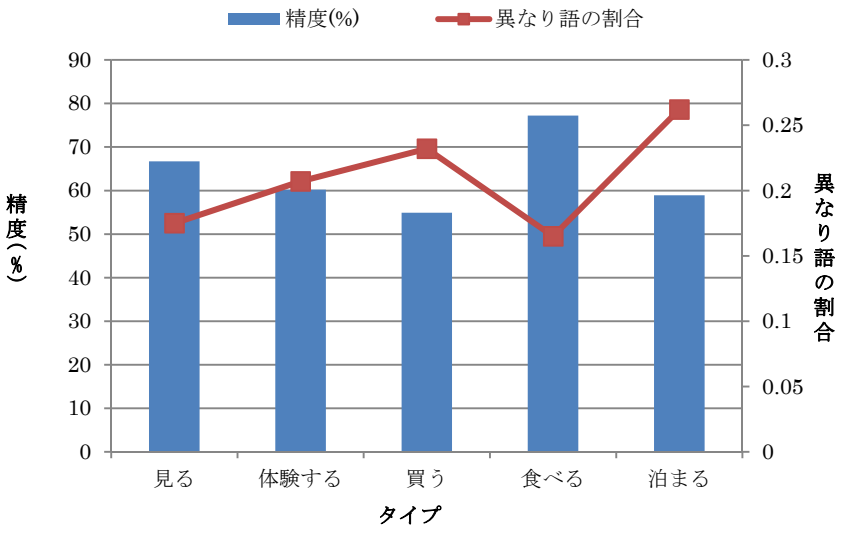


図 5.7　旅行ブログエントリのタイプ分類の精度と異なり語の割合

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 5.10 に示す．

表 5.10　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 5.10 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―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 5.6 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 5.11 にまとめる．表 5.11 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 5.11　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 5.11 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[43]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．

5.5.2　旅行ガイドブックのブック単位への対応付け

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．5.4.2 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 5.12，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 5.13 に示す．

表 5.12　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 5.13　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 5.12，表 5.13 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[44]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[45, 46]．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．

5.5.3　旅行ガイドブックのページ単位への対応付け

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，5.5.2 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，5.5.2 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 5.14 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 5.14　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

5.5.4　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 5.8 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 5.8 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．

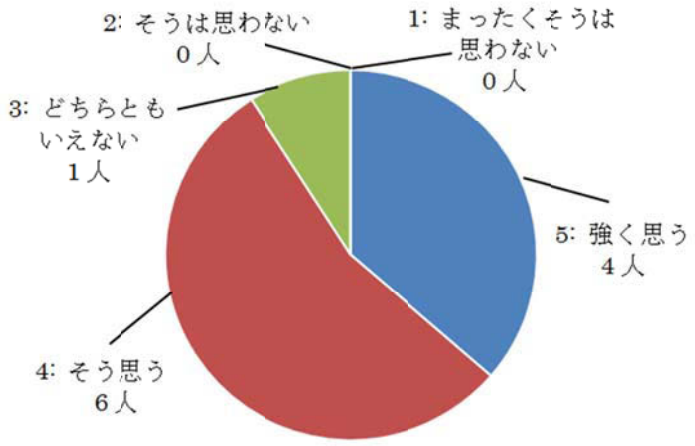


図 5.8　情報拡張された旅行ガイドブックを閲覧するシステムの有用性評価

被験者による自由記述を表 5.15 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[24, 25, 26]．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．

表 5.15　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．

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*16，質問応答コンテンツは Yahoo!知恵袋の検索システム*17（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 5.16 に示す．

表 5.16　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．

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

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 5.15 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 5.16 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．

特に，表 5.16 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．

5.6 まとめ

本章では，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を行う手法について述べた．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．

第6章　結論

本研究では，まず，観光情報の有用な情報源として旅行ブログエントリに着目し，ブログデータベースから旅行ブログエントリを自動で抽出する手法を提案した．旅行ブログエントリの抽出に関しては，精度 86.7%，再現率 38.1%を得た．また，旅行ブログエントリから，土産物情報，観光名所情報を抽出することで，旅行ブログエントリの観光情報の情報源としての有用性を確認した．

次に，自動で抽出した旅行ブログエントリを利用し，2 種類の観光支援システムを構築した．自動で抽出した旅行ブログエントリを利用することで，低コストで観光支援システムを作成することが可能になると考えられる．同時に，網羅性の高さや最新の観光情報を素早く獲得できる点などで，既存の観光を支援する媒体よりも有用なものになることが期待される．

旅行ブログエントリ中に含まれる観光情報を利用した観光支援システムとして，観光情報リンク集を構築した．旅行ブログエントリには，観光の際に参考にした Web ページへのリンクが，観光情報として提示されている．旅行ブログエントリ中に含まれるリンクを，観光情報リンクと呼ぶこととする．本研究では，旅行ブログエントリから観光情報リンクを収集し，「見る」，「食べる」などの観光に特化したタイプに分類することで，自動で観光情報リンク集を構築した．タイプ分類には，観光情報リンク周辺の文字列を使用した．実験の結果，精度 76.9%，再現率 66.4%を得た．

また，ソーシャルメディア上の投稿を，既存の観光情報データベースと組み合わせた観光支援システムとして，情報拡張した旅行ガイドブックの閲覧システムを構築した．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．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のか，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，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と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手法を提案する．実験の結果，旅行ガイドブックのページへ，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，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できるシステムの構築を行い，被験者による評価により，提案システムが旅行の計画を行う際に有用であることを示した．

本研究では，日本語で記述された旅行ブログエントリを対象としたが，英語で記述された旅行ブログエントリも対象とすることで，外国人旅行客の観光支援システムの構築への応用も可能である．また，本研究では主に，旅行ブログエントリに含まれるテキスト情報を使用して観光支援システムを構築している．今後の研究では，flickr などの画像共有サイトや YouTube などの動画共有サイトに投稿された写真や動画も対象とすることで，より視覚的な情報も提供できる観光支援システムを構築していく予定である．

謝辞

本研究を行うに当たり，格別なる御指導ならびに御鞭撻を賜りました竹澤寿幸教授，難波英嗣准教授に深甚なる感謝の意を表します．

北上始教授，高橋健一教授には，お忙しい中，博士論文の審査をお引き受けいただき，公聴会などで研究に関する助言をいただきました．深く御礼申し上げます．

相澤輝昭先生には，自然言語処理の基礎を教えていただき，様々な助言をいただきました．深く感謝致します．公聴会には，遠いところをご足労いただき，有難うございました．

楽天技術研究所ニューヨークの関根聡さん，村上浩司さんには，インターンシップで受け入れて頂き，手厚くご指導いただきました．大企業で研究をするという大変貴重な経験をさせていただきました．心より感謝申し上げます．

黒澤義明助教，目良和也助教には，日頃から有益なご助言をいただき，多面に渡って励ましていただきました．有難うございました．

本論文をまとめるに当たって御協力いただいた言語音声メディア工学研究室の皆様に厚く御礼申し上げます．皆様と一緒に，研究について語り合い，談笑した日々は大切な思い出です．

旅行ガイドブックの情報拡張の実験では，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いただきました．深く御礼申し上げます．

最後に，私を常に温かく見守り，支えてくれた家族に感謝致します．

参考文献

[1] 大槻洋輔, 佐藤理史, “地域情報ウェブディレクトリの自動編集”, 情報処理学会論文誌, Vol.42, No.9, pp.2310-2318, 2001.

[2] 佐藤理史, “ワールドワイドウェブを利用した住所検索”, 情報処理学会論文誌, Vol.42, No.1, pp.59-67, 2001.

[3] 村山紀文, 南野朋之, 奥村学, “メタデータ付与のための住所録自動生成”, 言語処理学会, 第 11 回年次大会, pp.53-56, 2005.

[4] Dmitry Davidov, “Geomining : Discovery of Road and Transport Networks Using Directional Patterns”, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp.267-175, 2009.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Hiroshi Kori, “Automatic Generation of Multimedia Tour Guide from Local Blogs”, Advances in Multimedia Modeling Lecture Notes in Computer Science Vol. 4351, pp. 690-699, 2006.

[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] 金子昂夢, 柳井啓司, “位置情報付き画像ツイートを利用した視覚的なイベント検出”, 第 5 回データ工学と情報マネジメントに関するフォーラム, 2013.

[12] 奥健太, 橋本拓也, 上野弘毅, 服部文夫, “位置情報付きツイート対応付けに基づく観光スポット推薦システムの開発”, ARG 第 2 回 Web インテリジェンスとインタラクション研究会, 2013.

[13] Yuki Arase, Xing Xie, Takahiro Hara, Shojiro Nishio, “Mining People’s Trips from Large Scale Geo-tagged Photos”, Proceedings of the international conference on Multimedia, pp.133-142, 2010.

[14] Tye Rattenbury, Nathaniel Good, Mor Naaman, “Towards automatic extraction of event and place semantics from flickr tags”, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp.103-110, 2007.

[15] 島田恵輔, 石野亜耶, 難波英嗣, 竹澤寿幸, “観光イベントに関する動画検索システムの開発”,日本データベース学会 第 5 回ソーシャルコンピューティングシンポジウム講演論文集, 2014.

[16] Fredric C. Gey, Ray R. Larson, Mark Sanderson, Hideo Joho, Paul Clough, Vivien Petras, “GeoCLEF: The CLEF 2005 Cross-Language Geographic Information Retrieval Track Overview”, Lecture Notes in Computer Science, LNCS4022, pp.908-919, 2005.

[17] Einat Amitay, Nadav Har'El, Ron Sivan, Aya Soffer, “Web-a-where: geotagging web content”, Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp.273-280, 2004

[18] Orkut Buyukkokten, Junghoo cho, Hector Garcia-molina, Luis Gravano, Narayanan Shivakumar, “Exploiting Geographical Location Information of Web Pages”, Procceedings of the ACM SIGMOD Workshop on the Web and Databases, pp.91-96, 1999.

[19] Rafael Odon de Alencar, Clodoveu Augusto Davis Jr., Marcos André Gonçalves, “Geographical classification of documents using evidence from Wikipedia”,

Proceedings of the 6th Workshop on Geographic Information Retrieval, Article No. 12, 2010.

[20] 安田宜仁, 戸田浩之, “検索位置のごく周辺を対象とした地理情報検索”, 人工知能学会論文誌, Vol.23, No.5C, pp.364-373, 2008.

[21] 相良毅, 喜連川優, “Web からの効率的な新規店舗の発見・登録支援手法”, 情報処理学会論文誌, Vol.48, No.SIG_11(TOD_34), pp.49-57, 2007.

[22] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, “Construction of a System for Providing Travel Information along Hiroden Streetcar Lines”, Proceedings of the 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM 2012), 2012.

[23] 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, 2014.

[24] Norihito Yasuda, Tsutomu Hirao, Jun Suzuki, Hideki Isozaki, “Identifying Bloggers’ Residential Areas”, Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, pp.231-236, 2006.

[25] Daisuke Ikeda, Hiroya Takamura, Manabu Okumura, “Semi-supervised Learning for Blog Classification”, Proceedings of the 23rd AAAI Conference on Artificial Intelligence, pp.1156-1161, 2008.

[26] Jonathan Schler, Moshe Koppel, Shlomo Argamon, James Pennebaker, “Effects of Age and Gender on Blogging”, Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs, pp.199-205, 2006.

[27] Daisuke Kawahara, Sadao Kurohashi, “A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis”, Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp.176-183, 2006.

[28] Anubhav Kale, Amit Karandikar, Pranam Kolari, Akshay Java Tim Finin, Anupam Joshi, “Modeling Trust and Influence in the Blogosphere Using Link Polarity”, International Conference on Weblogs and Social Media, 2007.

[29] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, “Automatic Classification of Link Polarity in Blog Entries”, Proceedings of the 7th Asian Information Retrieval Societies Conference (AIRS2011), 2011.

[30] Justin Martineau, Matthew Hurst, “Blog Link Classification”, Proceedings of International Conference on Weblogs and Social Media, 2008.

[31] 難波英嗣, “論文間の参照情報の抽出と利用に関する研究”, 北陸先端科学技術大学院大学 博士論文, 2001.

[32] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, Krishnaram Kenthapadi , “Enriching Textbooks with Images”, Proceedings of the 20th ACM Conference on Information and Knowledge Management, pp.1847-1856, 2011.

[33] Andrei Broder, Marcus Fontoura, Vanja Josifovski, Lance Riedel, “A Semantic Approach to Contextual Advertising”, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.559-566, 2007.

[34] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, “Providing Ad Links to Travel Blog Entries Based on Link Types”, Proceedings of the 9th Workshop on Asian Language Resources collocated with IJCNLP 2011, pp.63-70, 2011.

[35] 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”, 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.

[36] 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.

[37] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, Cédric Bray, “Visual Categorization with Bags of Keypoints”, Proceedings of ECCV International Workshop on Statistical Learning in Computer Vision, pp.1-22, 2004.

[38] 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.

[39] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, Chong-Wah Ngo, “Evaluating Bag-of-Visual-Word Representation in Scene Classification”, Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, pp.197-206, 2007.

[40] Qi He, Jian Pei, Simon Fraser, Daniel Kifer, Prasenjit Mitra, Lee Giles, “Context-aware Citation Recommendation”, Proceedings of the 19th international conference on World Wide Web, pp.421-430, 2010.

[41] Noriko Kando, “Text-level Structure of Research Papers: Implications for Text-Based Information Processing Systems”, Proceedings of the British Computer Society Annual Colloquium of Information Retrieval Research, pp.68-81, 1997.

[42] Gerard Salton, “The SMART Retrieval System –Experiments in Automatic Document Processing”, Prentice-Hall, Inc., Upper Saddle River, NJ, 1971.

[43] 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.

[44] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, “Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries”, Proceedings of the International Conference on Information Technology and Travel &Tourism, pp.113-124, 2011.

[45] 晃舜祥恵, 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.

[46] 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.

発表論文一覧

受賞

- 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 2014) 学生プレゼンテーション賞
  石野亜耶, 藤井一輝, 藤原泰士, 前田剛, 難波英嗣, 竹澤寿幸, “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”, 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.
- 最優秀修士論文賞
  石野亜耶, “ブログを中心とした観光情報の組織化”, 広島市立大学大学院 修士論文, 2011.

学位論文

- 石野亜耶, “ブログを中心とした観光情報の組織化”, 広島市立大学大学院 修士論文, 2011.
- 石野亜耶, “リンクの評価極性に基づいたブログサイトの質の評価”, 広島市立大学 学士論文, 2009.

論文誌

[1] 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.

[2] 石野亜耶, 難波英嗣, 竹澤寿幸, “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”, 人工知能学会論文誌, Vol.29, No.3, 2014. （2014 年 4 月 16 日採録決定）

国際会議

[1] Hidetsugu Nanba, Ryuta Saito, Aya Ishino, Toshiyuki Takezawa, “Automatic Extraction of Event Information from Newspaper Articles and Web Pages”, Proceedings of the 15th International Conference on Asia-Pacific Digital Librariesthe (ICADL 2013), LNCS 8279, pp.171-175, Bangalore, India, 2013.

[2] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, “Extracting Transportation

Information and Traffic Problems from Tweets During a Disaster: Where do you evacuate to?", Proceedings of the Second International Conference on Advances in Information Mining and Management (IMMM 2012), Venice, Italy, 2012.

[3] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, "Construction of a System for Providing Travel Information along Hiroden Streetcar Lines", Proceedings of the 3rd IIAI International Conference on e-Services and Knowledge Management (IIAI ESKM 2012), Fukuoka, Japan, 2012.

[4] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, "Automatic Classification of Link Polarity in Blog Entries", Proceedings of the 7th Asian Information Retrieval Societies Conference (AIRS2011), Dubai, United Arab Emirates, 2011.

[5] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, "Providing Ad Links to Travel Blog Entries Based on Link Types", Proceedings of the 9th Workshop on Asian Language Resources collocated with IJCNLP2011, Chiang Mai, Thailand, 2011.

[6] Aya Ishino, Hidetsugu Nanba, Toshiyuki Takezawa, "Automatic Compilation of an Online Travel Portal from Automatically Extracted Travel Blog Entries", Proceedings of the 18th international Conference on Information Technology and Travel &Tourism (ENTER2011), Innsbruck, Austria, 2011.

[7] Hidetsugu Nanba, Haruka Taguma, Takahiro Ozaki, Daisuke Kobayashi, Aya Ishino, Toshiyuki Takezawa, "Automatic Compilation of Travel Information from Automatically Identified Travel Blogs", Proceedings of the 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, Singapore, 2009.

その他の発表論文

[1] 石野亜耶, 村上浩司, 関根聡, "商品レビューからの購買意図の抽出とそれを用いた商品検索システムの構築", 言語処理学会 第 20 回年次大会, 2014.

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

[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] 島田恵輔, 山本夏生, 石野亜耶, 難波英嗣, 竹澤寿幸, “観光イベントに関する動画とブログの自動収集”, 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.

[5] 前田剛, 河野有希, 石野亜耶, 難波英嗣, 竹澤寿幸, “場所に焦点を当てた複数ブログの自動要約”, 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.

[6] 石野亜耶, 藤井一輝, 藤原泰士, 前田剛, 難波英嗣, 竹澤寿幸, “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”, 第 5 回 Web とデータベースに関するフォーラム(WebDB Forum 2012), 2012.

[7] 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(WebDB Forum 2012), 2012.

[8] 石野亜耶, 難波英嗣, 竹澤寿幸, “広電沿線観光情報提示システムの構築”, インタラクティブ情報アクセスと可視化マイニング研究 会キックオフ・イベント＆第 1 回研究会, 2012.

[9] 石野亜耶, 小田原周平, 難波英嗣, 竹澤寿幸, “Twitter からの被災時の行動経路の自動抽出および可視化”, 言語処理学会 第 18 回年次大会, 2012.

[10] 石野亜耶, 難波英嗣, 竹澤寿幸, “ブログ中のリンクの評価極性判定”, 言語処理学会 第 18 回年次大会, 2012.

[11] 寺西拓也, 野村達二, 平山智子, 石野亜耶, 難波英嗣, 竹澤寿幸, “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”, 言語処理学会 第 18 回年次大会, 2012.

[12] 石野亜耶, 難波英嗣, 竹澤寿幸, “Twitter からの危険情報の抽出”, 第 4 回楽天研究開発シンポジウム, 2011.

[13] 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.

[14] 小田原周平, 石野亜耶, 難波英嗣, 竹澤寿幸, “ブログからのユーザの行動経路の自動抽出と可視化”, 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.

[15] 石野亜耶, 難波英嗣, 竹澤寿幸, “ブログを中心とした観光情報の組織化”, 第 3 回楽天研究開発シンポジウム, 2010.

[16] 石野亜耶, 小林大祐, 難波英嗣, 竹澤寿幸, “ブログを利用した観光情報リンク集の自動構築”, 言語処理学会 第 16 回年次大会, 2010.

[17] 石野亜耶, 難波英嗣, 田熊遥, 尾崎貴紘, 小林大祐, 竹澤寿幸, “旅行ブログからの観光情報の自動抽出”, 電子情報通信学会第 15 回 Web インテリジェンスとインタラクション研究会, pp.19-23, 2009.