

Wikipediaからの地名語句抽出に基づく ジオタグベースクロスドメイン推薦システム

那須 昇平[†] 奥 健太[†]

[†] 龍谷大学理工学部 〒520-2194 滋賀県大津市瀬田大江町横谷 1-5

E-mail: [†]t150520@mail.ryukoku.ac.jp, ^{††}okukenta@rins.ryukoku.ac.jp

あらまし クロスドメイン推薦システムは映画やコミック、ゲームなど異種のジャンルにまたがってアイテムを推薦システムである。本研究ではアイテムに付与された位置情報を基にドメインをまたがって関連するアイテムを推薦するシステムを提案する。まず、アイテムの wikipedia 記事のストーリー部分やあらすじ部分のから形態素解析により地名語句を抽出する。ジオコーディングを用いてその語句の位置情報を取得することで、アイテムに位置情報を付与する。位置情報を付与したアイテムに基づき、入力地域から関連アイテムを推薦する地域-アイテム推薦方式、入力アイテムから関連地域を推薦するアイテム-地域推薦方式を提案した。提案システムの振舞いを分析するために、地域-アイテム推薦およびアイテム-地域推薦のそれぞれの結果を示し、考察した。提案システムにより、地域に関連する作品や、作品に関連する地域が推薦されることを示した。一方で、いくつかの課題もみられたため、アイテム被覆率および関連地域の再現性、地名の曖昧性の観点から提案システムの限界について考察した。

キーワード クロスドメイン推薦, 位置情報, Wikipedia

1. はじめに

2017 年の訪日外客数が 2,800 万人に達する^(注1)など観光を楽しむ人々は年々増大してきている。自然や街並みなどの景観を眺めたり、美術館や博物館巡りを楽しんだり、現地の名物料理を食したりするなど、観光の楽しみ方はさまざまである。また、映画や小説等の作品の舞台となった地域へ訪れてみるということも観光を楽しむ方法の一つである。さらには、観光体験を通じて、観光地に関連する作品と出会うことも、自身の興味を拓けることの契機になる。

このような背景の下、我々の先行研究[1][2]では、位置情報ベースのクロスドメイン推薦システムを提案してきた。クロスドメイン推薦システムは、書籍や映画、楽曲など、複数の異種ドメインにまたがってアイテムを推薦するシステムである。クロスドメイン推薦を実現するためには、複数のドメインにおける共通属性を抽出する必要がある。我々は、この共通属性として位置情報に着目している[1][2]。各ドメインのアイテムと位置情報とを関連付けておくことで、次の二つの推薦機能を実現している；(a) 位置情報-アイテム推薦：入力された位置情報に関連するアイテムの推薦、(b) アイテム-位置情報推薦：入力されたアイテムに関連する位置情報を推薦。先行研究[1][2]では、アイテムと位置情報の関係データをクラウドソーシングにより収集した。

本研究では、クラウドソーシングに依存しない方法として、Wikipedia^(注2)からアイテムに関する位置情報を抽出する方式を検討する。本研究では、映画や漫画、ゲームなど位置情報と

関連付けられやすいドメインを対象とする。また、これらのドメインに含まれるアイテム（作品）のうち、Wikipedia に登録されているアイテムを対象とする。各アイテムの Wikipedia 記事の「ストーリー」や「あらすじ」部分から抽出した地名語句を基に、各アイテムに位置情報を付与する。本研究では、Wikipedia を基に関連付けられた位置情報をベースとしたクロスドメイン推薦システムを提案する。

なお、推薦システムの基本方式として、大きく次の三つの方式がある；(a) 協調ベース推薦[3]、(b) 内容ベース推薦[4]、(c) 知識ベース推薦[5][6]。(a) および (b) の方式には、大量の履歴データが必要となる。また、(a) では新規ユーザおよび新規アイテムのコールドスタート問題がある。(b) では、新規アイテムのコールドスタート問題については解消されるものの、新規ユーザのコールドスタート問題は依然として残る。これらに対し、(c) は履歴データは必要とせず、新規ユーザに対しても推薦可能である。

知識ベース推薦には、制約ベース推薦方式と事例ベース推薦方式がある。制約ベース推薦は、ユーザが提示した制約を満たすアイテムを推薦する。事例ベース推薦は、ユーザが例示したアイテムや属性に類似するアイテムを推薦する。推薦されたアイテムを参考にしながら、ユーザが条件を修正（批評）していくことで、ユーザのアイテム探索を支援する。本研究では、履歴データをもたない新規ユーザにも推薦可能なシステムを想定しているため、知識ベース推薦方式に着目する。特に、ユーザが興味アイテムや属性（位置情報）をシステムに入力することで、類似するアイテムを推薦するため、本研究で提案するシステムは事例ベース推薦方式に位置付けられる。

事例ベース推薦方式により推薦システムを実装する際には、アイテム間の類似性尺度の定義と批評インタフェースの設計が

(注1) : <https://statistics.jnto.go.jp/>

(注2) : <https://ja.wikipedia.org/wiki/>

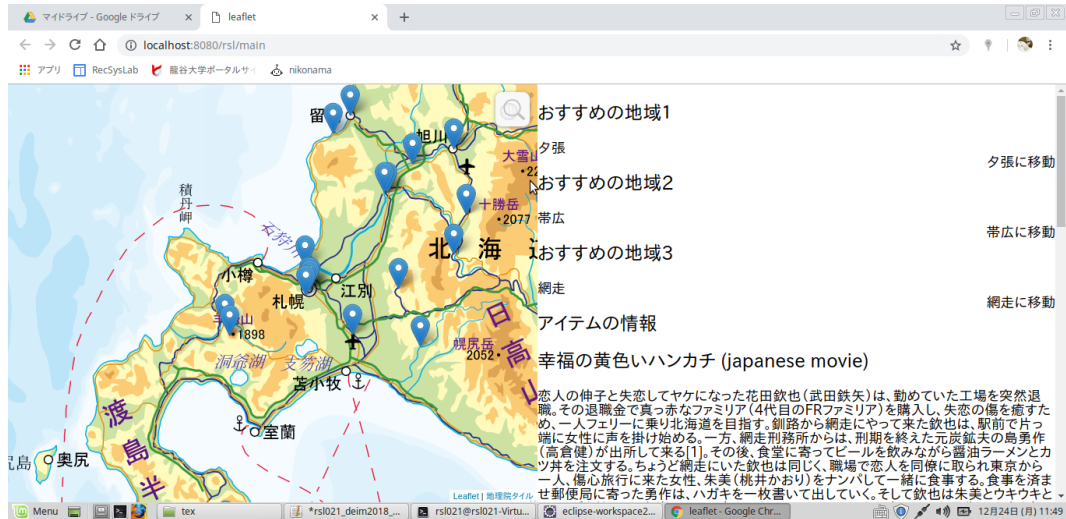


図 1 ジオタグベースクロスドメイン推薦システムのインターフェース。地図画像は国土地理院の電子地形図（タイル）を掲載。

必要となる。本研究では、アイテムの共通属性である位置情報を基にアイテム間に関連付ける。また、位置情報に関連付けられたアイテムを可視化するための地図インターフェースを設計する。ユーザは地図操作を行うことで、位置情報や興味アイテムを修正することができる。地図インターフェースを介して、更新された位置情報に関連するアイテムの推薦—位置情報-アイテム推薦—、選択したアイテムに関連する位置情報の推薦—アイテム-位置情報推薦—の各機能を提供する。

2. 関連研究

2.1 クロスドメイン推薦システム

従来の推薦システム [5] は単一のドメインにおいて嗜好データを獲得し、それに基づき当該ドメインにおける推薦を行うという方法が主流であった。例えば、映画推薦システムでは、映画の視聴履歴や評価履歴に基づき対象ユーザの嗜好に関する学習モデルを獲得する。その学習モデルを用いてやはり映画ドメインにおける嗜好を予測する。しかし、単一ドメインにおいて獲得された学習モデルは他のドメインにおいては利用できない。仮に映画ドメインにおいて十分な嗜好データが蓄積されていたとしても、それをそのまま書籍ドメインにおける嗜好の予測に用いることはできない。

そこで、近年クロスドメイン推薦という考え方が注目されてきた [7]。クロスドメイン推薦システムは、複数のドメインにまたがった推薦を実行するシステムである。例えば、映画ドメインで蓄積された嗜好データを基に、書籍ドメインにおける嗜好を予測するというものである。これは異なるドメインにおいてもドメイン間の嗜好には何らかの関連性があるという考えに基づく。

我々は、複数のドメインにおける共通属性として位置情報に着目し、これまでに位置情報を介したクロスドメイン推薦システムを提案してきた [1] [2] これら先行研究 [1] [2] では、アイテムと位置情報の関係データをクラウドソーシングにより収集した。これらに対し、本研究では、クラウドソーシングに依存し

ない方法として、Wikipedia からアイテムに関する位置情報を抽出する方式を検討する。各アイテムの Wikipedia 記事の「ストーリー」や「あらすじ」部分から抽出した地名語句を基に、各アイテムに位置情報を付与する。

3. 定義

定義 1：アイテム。 映画や漫画、ゲームなどの作品をアイテムとよび $a_i \in A$ と表す。本稿では、アイテム種別として、映画作品、日本の映画作品、日本のアニメ映画作品、日本の漫画作品、コンピュータゲームの 5 種類のアイテムを扱う。アイテム a_i のタイトルを $a_i.title$ とする。また、アイテム a_i に関する Wikipedia の記事に含まれるテキストを $a_i.text$ とする。

定義 2：地名語句集合。 すべての地名語句集合を G と表す。アイテム a_i に関連付けられた地名語句集合を $G^i \in G$ と表す。アイテム a_i に関連付けられた地名語句 $g_k^i \in G^i$ は、その地名に対応する地理座標系の位置座標として経度・緯度を $(g_k^i.x, g_k^i.y)$ をもつ。

4. システム

本章では、提案システムであるジオタグベースクロスドメイン推薦システムについて説明する。

4.1 システム概要

クロスドメイン推薦システムは映画やコミックやゲームなど異種のジャンルにまたがってアイテムを推薦するシステムである。ジオタグベースクロスドメイン推薦システムは、ユーザが地図操作を行う過程で、興味地域に関連するアイテムや、さらに関連する他の地域を提示することで、アイテムや地域の発見的な探索を可能とするシステムである。特に、アイテムの種別としては、一つのドメインに限定せず、異種ドメインのアイテムを扱うため、地域を介した異種のドメインにまたがったアイテムの探索を可能とする。なお、本稿では、映画作品、日本の映画作品、日本のアニメ映画作品、日本の漫画作品、コンピュータゲームの 5 種類をアイテムのドメインとして扱っている。

ストーリー [編集]

「名探偵コナンの事件」も参照

高校生探偵・**工藤新一**（くどう しんいち）は大人顔負けの推理力と洞察力を併せ持ち、**「日本警察の救世主」**とまで言われていた。

ある日、新一は幼なじみの**毛利蘭**（もうり らん）と遊園地遊びに行くが^[注 5]、そこで謎の組織の取引現場を目撃する。取引の様様を見るのに夢中になっていた新一は、組織の持ち主1人の仲間によって背後から殴り倒され、口封じのために毒薬を飲まされる。だが、組織も知らなかった薬の副作用によって神経を除いたすべての組織が退行し、小学1年生の体となってしまう。

新一は、自分が生きていることを組織に知られたら再び命を狙われ、周囲の人間も巻き添えにしようとする隣人の**阿笠博士**（あがさはかせ）から助言を受ける。そして、正体を隠し**江戸川コナン**（えどがわ コナン）と名乗り^[注 6]、蘭の父・**毛利小五郎**（もうり こごろう）が探偵であることから謎の組織の情報が得られると考え、蘭の家に居候する。

以後、コナンは周囲で次々に起きる事件を持ち前の推理力や、阿笠の開発した万能アイテムを駆使して解決していく。同時に、元の身体を取り戻すために謎に包まれた**黒の組織**の陰謀を追っていくことになる。

登場人物 [編集]

「名探偵コナンの登場人物」を参照

江戸川 コナン（えどがわ コナン） / **工藤 新一**（くどう しんいち）

本作の主人公。高校生探偵として名を響かせていたが、**黒の組織**によって試作段階の毒薬（**APTX4869**）を飲まされ、身体が縮んで幼児化してしまった。組織に新一の生存がバレて周囲に危害が及ぶのを防ぐため、事情を知る協力者たちの力を借りて、普通の小学生として生活しながら事件を解決していく。サッカーの実力は相当なものである。

毛利 蘭（もうり らん）

本作のヒロイン。コナンのお姉さん的な存在で、コナンを弟のようにかわいがっている。また、幼なじみの新一の帰りを待ちながら、新一との恋愛関係も少しずつ進展している。空手の達人でもある。

毛利 小五郎（もうり こごろう）

図 2 Wikipedia の『名探偵コナン』のストーリー・登場人物欄 (<https://ja.wikipedia.org/wiki/名探偵コナン>) .

図 1 は本システムのインターフェースである。インターフェースはマップビューと出力ビューから構成される。ユーザはマップビューから興味のある地域を地図操作を行うことで閲覧することができる。地図操作中に、その地図上に表示されている領域に関連するアイテムが提示される。さらに、ユーザが興味のあるアイテムのアイコンをクリックすると、出力ビューにアイテムの詳細情報および関連地域が提示される。関連地域とは、そのアイテムに関して、現在表示されている領域以外に関連する地域を指す。ユーザはこの操作を繰り返すことで、興味地域から関連アイテムの探索、興味アイテムから関連地域の探索を行うことができ、アイテムや地域の発見的な探索が可能となる。

このようなシステムを実現するためには、あらかじめアイテムに関して位置情報を付与しておく必要がある。本稿では、アイテムへの位置情報付与をアイテムへのジオタギングとよぶ。4.2 節では、アイテムへのジオタギングについて説明する。さらに、4.3 節では、地域-アイテム推薦およびアイテム-地域推薦の各手法について説明する。

4.2 アイテムへのジオタギング

4.2.1 アイテムに関するテキストの抽出.

各アイテム種別について、Wikipedia の一覧ページ（例：日本の映画作品の場合は「日本の映画作品一覧」ページ^(注3)）からアイテムリスト A を抽出する。各アイテム $a_i \in A$ について、そのアイテムの記事ページから、「ストーリー」や「あらすじ」部分のテキストを抽出し、そのテキストを $a_i.\text{text}$ とする。ここで、後述するアイテムへのジオタギングにおいては、アイテムの内容に基づき関連する位置情報を付与する。これら、「ストーリー」や「あらすじ」には、アイテムの内容に関する記述が含まれるため、本稿ではこれらの部分をアイテムの位置情報を抽出する対象としている。

4.2.2 地名語句の抽出

抽出した $a_i.\text{text}$ を対象に形態素解析を行う。本研究では、形態素解析器として MeCab を用いる。MeCab によって品詞が地名語句として抽出された語句集合をそのアイテムに関連付けらえた地名語句集合 G^i とする。抽出されたすべての地名語句集合を G とする。

ただし、抜き出された地名語句の中に、当該アイテムに登場する人物の名前が地名として抽出されることがある。この問題を解消するために、事前に当該アイテムの記事ページ中の「登場人物」の部分から人物名リストを取得しておく。その後、地名語句集合にその人物名が含まれていれば、それを削除する。

図 2 には、例として Wikipedia の名探偵コナンの登場人物欄を示している。ストーリー欄から抜き出される地名語句は「日本」、「毛利」、「江戸川」の三つであるが、「毛利」および「江戸川」は登場人物名であり地名語句ではない。これらの語句を取り除くために、登場人物の中から「江戸川」、「コナン」、「工藤」、「新一」、「毛利」、「蘭」などの登場人物一覧を作成しておく。抽出された地名語句集合から、これら人物名を表す語句を削除する。この例では、結果として残る地名語句は「日本」のみということになる。

4.2.3 地名語句への経度・緯度の付与

地名語句 $g_k^i \in G$ について、経度・緯度 $(g_k^i.x, g_k^i.y)$ を付与する。経度・緯度の付与にはジオコーディングを用いる。

4.3 推 薦

4.3.1 地域-アイテム推薦

入力された経度・緯度 (x, y) に基づき、周辺の関連するアイテム集合を地図上に提示する。具体的には、地図の表示領域を取得し、その領域内に含まれる地名語句集合を取得する。その地名語句集合に関連するアイテム集合を地図上に提示する。

4.3.2 アイテム-地域推薦

入力されたアイテム a_i に基づき、他に関連する地域を提示

(注3) : <https://ja.wikipedia.org/wiki/日本の映画作品一覧>

する．具体的には，アイテム a_i に関連付けられた地名語句集合 G^i を取得し，その地名語句 $g_k^i \in G^i$ を提示する．

5. 結 果

提案システムによる出力結果を踏まえ，システムの振舞いを考察する．まず，今回用いたデータセットについて述べた後，地域-アイテム推薦（5.2 節）およびアイテム-地域推薦（5.3 節）のそれぞれの結果を示す．

5.1 データセット

本研究では，アイテム種別として，映画作品，日本の映画作品，日本のアニメ映画作品，日本の漫画作品，コンピュータゲームの 5 種類のアイテムを対象としている．評価用データセットとして，Wikipedia の各ドメインに対応する一覧ページにリンクが含まれる記事を対象に収集した．表 1 にドメインごとのデータの概要を示す．表 1 には，情報源とした Wikipedia 記事を示している．また，各ドメインに含まれるアイテム数およびジオタグアイテム数，その割合を示している．

5.2 地域-アイテム推薦結果

地域-アイテム推薦結果について考察する．ここでは，「甲子園」および「沖縄」，「イギリス」の三つの地域を入力したとき，それぞれシステムによる出力結果について考察する．

a) 甲 子 園

入力経度・緯度を「甲子園」(135.361637, 34.721355) としたとき，提示されたアイテム集合について分析した．図 3 は，「甲子園」を入力地域としたときの提示されたアイテム集合を示す．図のとおり，『タッチ 3』や『私は野球部マネージャー』，『剛球少女』，『もし高校野球の女子マネージャーがドラッカーの『マネジメント』を読んだら』などが提示された．このように，甲子園に関連する多くの漫画作品が提示された．

b) 沖 縄

入力経度・緯度を「沖縄」(127.80558, 26.334427) としたとき，提示されたアイテム集合について分析した．『トップをねらえ!』や『アイカツ!』，『FULL』などが提示された．それぞれの作品で「沖縄」は，主人公の出身地であったり，コンサートの会場となっていたりするなど，提示されたすべての作品は沖縄に関連していた．このように，沖縄に関連する多くの映画作品や漫画作品が提示された．

c) イ ギ リ ス

入力経度・緯度を「イギリス」(55.378051, -3.435973) としたとき，提示されたアイテム集合について分析した．『スチームボーイ』や『ゴーストリコン 2』，『屍者の帝国』などが提示された．このうち，『スチームボーイ』，『屍者の帝国』，『思い出のマーニー』，『GIANT』，『レディ!!』については，イギリスが舞台の映画や，作品中にイギリスに訪れる機会のある作品であった．一方で，『ゴーストリコン 2』については，「イギリス，フランス，ドイツ，オーストラリアに支持されたアメリカ」といったようにアメリカ陣営に所属する国の一つとしてイギリスが紹介されたため，イギリスに関連のある作品とはいえない．このように，イギリスに関連する多くの映画作品やコンピュータゲーム，漫画作品が提示されたが，その一方で，イギリスに関

連の低いアイテムも提示されるという課題も見つかった．

5.3 アイテム-地域推薦結果

アイテム-地域推薦結果について考察する．ここでは，『君の名は。』および『ちはやふる』の二つの作品を入力したとき，それぞれシステムによる出力結果について考察する．

a) 『君の名は。』

日本の映画作品の一つである『君の名は。』を入力したとき，提示された地域について分析した．図 4 は『君の名は。』を入力したときに提示された地域を示す．図のとおり，関連地域として「岐阜」，「東京」，「飛騨」，「四ツ谷」，「新宿」などが提示された．これらは Wikipedia の記事から抽出された地名語句であり，関連地域として適切に選ばれていることがわかる．一方で，本作品にはモデルとなった実在の地域が多く登場する．例えば，「飛騨古川駅」や「気多若宮神社」などが挙げられる．しかし，これらの地域は本作品の中では架空の名称として描写されているため，Wikipedia の「ストーリー」や「あらすじ」には実在の地名としては明示的には含まれていない．このように，Wikipedia の「ストーリー」や「あらすじ」を情報源とした本システムでは，作品のモデルとなった地域の抽出は難しい．

b) 『ちはやふる』

日本の漫画作品である『ちはやふる』を入力したとき，提示された地域について分析する．「吉野」，「富士崎」，「北海道」，「明石」，「福井」，「東京」，「府中」が提示された．これらは Wikipedia の記事から抽出された地名語句であり，関連地域として適切に選ばれていることがわかる．具体的には，対戦相手の学校名（またはその一部）や競技かるたの大会名が多く抽出され，関連のある地名が多く抽出できた．一方で，本作品には「吉野会大会」という語が出現するが，本システムでは誤って奈良の「吉野」に関連付けられた．この問題については，次節で考察する．

5.4 限 界

5.2 節および 5.3 節では，それぞれの推薦結果について示し，地域に関連する作品や，作品に関連する地域が推薦されることを示した．一方で，いくつかの課題もみつかった．本節では，提案システムの限界について，アイテム被覆率，関連地域の再現性，地名の曖昧性の観点から考察する．

a) アイテム被覆率

表 1 に示すとおり，今回の評価実験では，合計で 9,966 件の作品を対象としたが，最終的に位置情報が付与された作品数は，全体の 17% にあたる 1,670 件に留まった．まず，作品を現実の地域と対応付けるという特性上，架空の世界を舞台とした作品は現実の地域とは関連付けられにくい．さらに，現実の地名語句が抽出されたアイテムにおいても，ジオコーディング手法の精度により，地名語句への位置情報付与に失敗した例も多い．このようなことから，アイテム全体に対する被覆率が低くなった．

b) 関連地域の再現性

本研究では，情報源として Wikipedia の「ストーリー」および「あらすじ」を対象とした．しかし，これらはあくまでも作品のあらすじを記述したものであり，作品中に出現する地域す

表 1 本実験で対象としたデータセット. ドメインごとに情報源とした Wikipedia 記事を示す.
また, 各ドメインに含まれるアイテム数およびジオタグアイテム数, その割合を示す.

ドメイン	情報源とした Wikipedia 記事	アイテム数	ジオタグアイテム数	割合
映画作品	「映画作品一覧」	1,813	553	30.5%
日本の映画作品	「日本の映画作品一覧」	844	258	30.6%
日本のアニメ映画作品	「日本のアニメ映画作品一覧」	1,218	124	10.2%
日本の漫画作品	「日本の漫画作品一覧」	4,556	704	15.5%
コンピュータゲーム	「コンピュータゲーム一覧」	1,535	51	3.3%
合計		9,966	1,690	17.0%

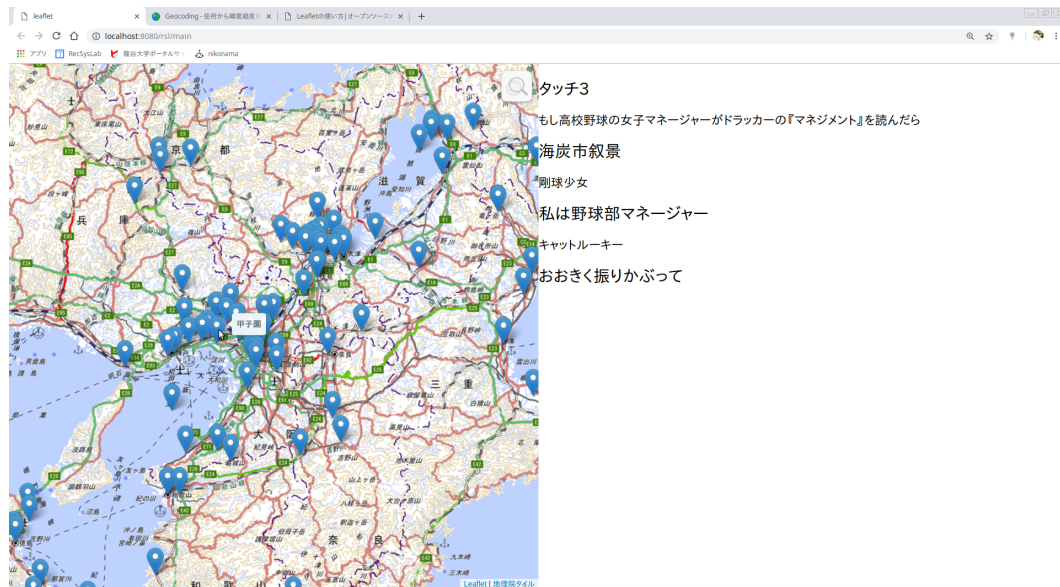


図 3 「甲子園」を入力地域としたときの出力画面. 地図画像は国土地理院の電子地形図（タイル）を掲載.



図 4 「君の名は。」を入力した時の推薦スポット. 地図画像は国土地理院の電子地形図（タイル）を掲載.

べてを網羅的に含むわけではない. また, 先述したとおり, 作品のモデルとして実在の地域が存在するものの, 作品中では地名が明示的には現れない, あるいは架空の名称が用いられてい

るような場合には, このような地域を「ストーリー」や「あらすじ」から抽出することは不可能である. このように「ストーリー」や「あらすじ」には含まれないが, 作品には関連するよ

うな地名を抽出するためには、別のアプローチを検討する必要がある。例えば、我々の先行研究であるクラウドソーシングによるアプローチ[1][2]と本システムを組み合わせるなどの方法が挙げられる。

c) 地名の曖昧性

Wikipedia を情報源とした本システムでは地名の曖昧性の問題が発生する。例えば、5.3 節の『ちはやふる』の結果で示した「吉野」の例のように、同一の地名であっても、本来の地域とは異なる地域として関連付けられることがある。このような地名の曖昧性を解消することが課題として残る。

6. おわりに

本稿では、Wikipedia からの地名語句抽出に基づくジオタグベースクロスドメイン推薦システムとして、地図インタフェースを介して位置情報に関連するアイテムの推薦-位置情報-アイテム推薦、選択したアイテムに関連する情報の推薦-アイテム-位置情報推薦の各機能について提供するシステムを提案した。

また、提案システムの振舞いを分析するために、地域-アイテム推薦およびアイテム-地域推薦のそれぞれの結果を示し、考察した。提案システムにより、地域に関連する作品や、作品に関連する地域が推薦されることを示した。一方で、いくつかの課題もみられたため、アイテム被覆率および関連地域の再現性、地名の曖昧性の観点から提案システムの限界について考察した。

謝 辞

本研究は JSPS 科研費 JP16H05932 の助成を受けたものです。ここに記して謝意を表します。

文 献

- [1] 北村統太, 奥健太. アイテムのジオタギングに基づく観光スポットベースクロスドメイン推薦システム. In *DEIM Forum 2017*, pp. B2-2, 2017.
- [2] 井田彰, 奥健太. ジオタグベースモーション順序を用いた連想型クロスドメイン推薦システム. In *DEIM Forum 2018*, pp. P6-4, 2018.
- [3] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, Vol. 22, No. 1, pp. 5-53, 2004.
- [4] Iván Cantador, Alejandro Bellogín, and David Vallet. Content-based recommendation in social tagging systems. In *[RecSys2010]Proceedings of the 4th ACM conference on Recommender systems*, p. 237, 2010.
- [5] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor. *Recommender Systems Handbook*. Springer, 2011.
- [6] Marius Kaminskis, Ignacio Fernández-Tobías, Francesco Ricci, and Iván Cantador. Knowledge-based music retrieval for places of interest. In *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies - MIRUM '12*, pp. 19-24, New York, New York, USA, 2012. ACM Press.
- [7] Ignacio Fernández-tobías, Iván Cantador, Marius Kaminskis, and Francesco Ricci. Cross-domain recommender systems : A survey of the State of the Art. In *Proceedings of the 2nd Spanish Conference on Information Retrieval2*, pp. 187-198, 2012.