

領域オントロジーと Linked Data を利用した観光情報推薦

Tourism Recommendation based on Ontologies and Linked Data

保科 宗淳
Muneaki HOSHINA

大河原 渉
Wataru OKAWARA

山口 高平
Takahira YAMAGUCHI

慶應義塾大学
Keio University

This paper discusses how application services go well with linked data in the field of recommendation systems for tourists. We produce linked data from related Web pages in order to get demographics such as age and gender, psychographics such as hobby and preference, behaviors such as life log, context such as time and place. We also discuss how to consume the linked data with domain ontologies in order to give useful recommendation system to the tourists.

1. はじめに

Linked Open Data(LOD)は、複数のオープンデータを RDF で表現し、相互にリンクづけられたデータ(Linked Data)であり、他のドメインデータへの連携がしやすいため、サービス機能の拡張が容易である(拡張容易性)。さらに、オントロジーとの連携により意味レベルの処理を行うことも可能である(高度化容易性)。しかし、Linked Data が急増する一方で、Linked Data の高度な利用についてはあまり考察されておらず、Linked Data のキラーアプリケーションは存在しないのが現状である。

このような背景から、本稿では情報推薦というドメインに着目し、(1) ユーザの年齢・性別といった「デモグラフィック属性」(2) 趣味や嗜好といった「サイコグラフィック属性」(3) 過去の行動履歴を表す「ビヘイビア属性」(4) 観光する月、エリア、目的、同行者といった「コンテキスト属性」という4種類のユーザセグメント情報を Linked Data として抽出し、オントロジーを利用することで、ユーザの嗜好や状況との適合性の高い観光スポットを推薦するシステムの構築方法について検討する。

2. 提案システム

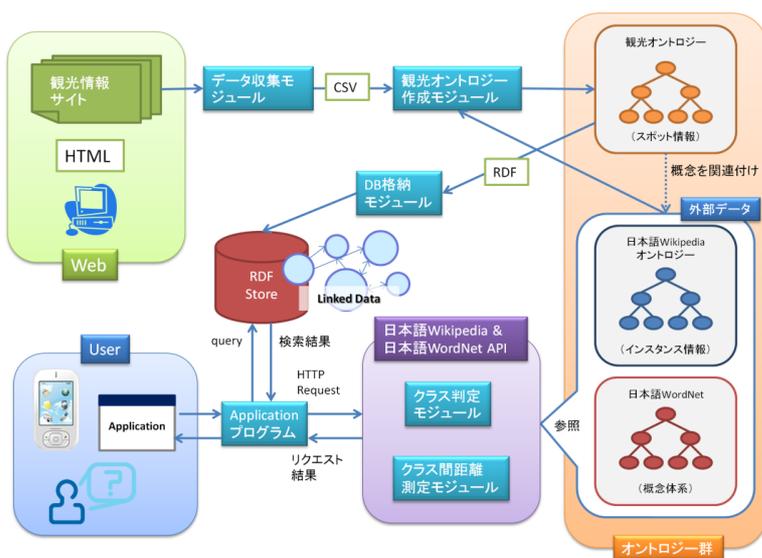


図 1: システム構成図

2.1 システム概要

2.1.1 データ収集および静岡観光オントロジーの作成

(1) データ収集モジュール

データ収集モジュールはじゃらん観光ガイド*1において静岡県の観光スポットとして登録されている全スポット(2011年12月現在 5041件)のうち、ユーザのクチコミ投稿が存在したスポットについて、スクレイピングによるデータ収集を行う。モジュールプログラミング言語 Python を用いて実装し、収集したデータは CSV 形式に変換して保存される。

(2) 静岡観光オントロジー作成モジュール

静岡観光オントロジー作成モジュールは、次の3種類のデータを合成し、一つのオントロジーファイル(静岡観光オントロジー: RDF形式)として出力するモジュールである。

- Protégé*2を用いて作成した、観光概念のクラス階層を記した RDFS ファイル(RDF形式)
- 観光概念と日本語 Wikipedia オントロジーおよび日本語 WordNet の関連付けを記した CSV ファイル

2.1.2 RDF データベース

DB格納モジュールは、PHP用のRDF & SPARQLライブラリであるARC2*3を用いて、静岡観光オントロジー作成モジュールで作成した静岡観光オントロジーファイル(RDF形式)をSPARQL検索が可能な形式でMySQLに保存する役割を果たしている。

なお、一度MySQLにデータを格納した静岡観光オントロジーファイルは、以後アプリケーションで直接利用することはない。格納された静岡観光オントロジーのデータはMySQLにARC2を通してSPARQLクエリを投げることで利用ができ、本システムのアプリケーションにおいてはそのような形のデータのやり取りを行っている。

2.1.3 アプリケーション

本システムのアプリケーションはスマートフォン向けのWebアプリケーションとしてPHPやjQuery Mobile等を用いて実装した。ユーザがアプリケーションで入力した情報を、MySQLに格納した静岡観光オントロジーのデータや日本語Wikipediaオントロジー[玉川11]と日本語WordNetを用いて分析し、そのユーザに適したおすすめスポットを提示する。

連絡先: 大河原渉, 山口高平
慶應義塾大学理工学部管理工学科
〒223-8522 神奈川県横浜市港北区日吉 3-14-1
TEL:045-566-1614
E-mail: w_okw@yahoo.co.jp, yamaguti@ae.keio.ac.jp

*1 <http://www.jalan.net/kankou/>
*2 <http://protege.stanford.edu/>
*3 <https://github.com/semsol/arc2/wiki>

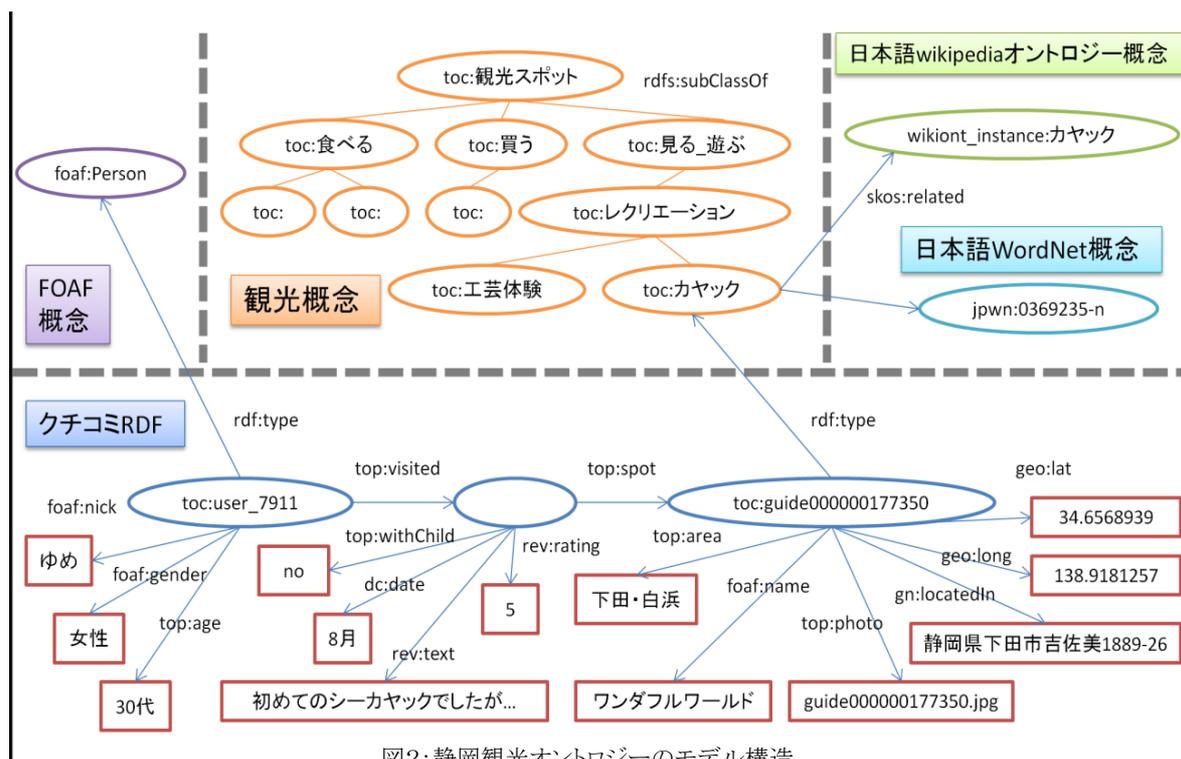


図2: 静岡観光オントロジーのモデル構造

2.2 静岡観光オントロジーのモデル構造

表1: 使用した標準語彙および個別語彙一覧

語彙名	説明
skos	上位下位関係など概念間の関係を記述するための語彙
dc	題名, 作者, 日付, 主題など, 一般的なデータ属性の語彙
rev	評価値や評価コメントなど評価について記述するための語彙
foaf	人々, 人々の活動, 人々間の関係を記述するための語彙
geo	緯度や経度など位置情報を記述するための語彙
gn	地理情報について記述するための語彙
toc※	独自に定義した観光に関するクラスを記述するための語彙
top※	独自に定義した観光に関するプロパティを記述するための語彙

本システムでは最終的にユーザに具体的な観光スポットを推薦する。そのために、具体的な観光スポットについての情報を標準語彙および独自に定義した個別語彙を利用して体系化したものが静岡観光オントロジーであり、図2にその一部を示す。本研究では観光スポットの情報を体系化する対象地域を静岡県に限定した。構築した静岡観光オントロジーの概要を以下に示す。

静岡観光オントロジーは大きく分けて、
 (1) 観光概念 (クラス階層)
 (2) クチコミRDF (インスタンスネットワーク)
 という二つの要素で構成されている。

(1) 観光概念

観光スポットの「種類」を列挙し、それらを事前に定めた弁別属性に従い分類・体系化したクラス階層である。弁別属性は、次の通りとなっている。

1層目はスポットの利用目的による分類である。観光における目的の分類方法は多々あると思われるが、本研究においては分類方法の一つとして観光サイトじゃらんによって「見る_遊ぶ」「買う」「食べる」の3種類に振り分けた。

- 「見る_遊ぶ」
食事や買い物を主目的としない観光スポットが該当する。下位クラスとして「レクリエーション」「文化景観」「自然景観」を持つ
- 「買う」
主に買い物を目的とするスポット。特に、買った物をその場で消費しない場合該当する。下位クラスとして「工芸品」、「農作物」、「醸造物」などを持つ。
- 「食べる」
食事に関するスポット。特に買った食べ物をその場で消費する場合該当する。下位クラスとして、「料理店」、「軽食」を持つ。

2層目は統一した分類方法ではなく、1層目の目的ごとに弁別属性を設定した。

- 1層目が「見る_遊ぶ」クラスの場合
行動による分類を行った。具体的には、博物館や街並みなど、「無生物を見る」ことを主目的とするスポットは「文化景観」に、植物や水族館など、「生物を見る」ことを主目的とするスポットは「自然景観」に、それ以外の遊びを主目的とするスポットは、「レクリエーション」にクラスを振り分けた。
- 1層目が「買う」クラスの場合
目的物による分類を行った。下位クラスとして、「工芸品」、「水産物」、「衣料物」、「農産物」、「醸造物」クラスを設定した。
- 1層目が「食べる」クラスの場合
店舗形式による分類を行った。お酒を主目的とした店は「バー_居酒屋」、レストランや料亭などは「料理店」、カフェやスイーツ店などは「軽食」とした。

(2)クチコミRDF (インスタンスネットワーク)

クチコミ RDF は、じゃらん観光ガイドに掲載されているクチコミ情報をスクレイピングし、表1のように標準語彙や独自に定めた個別語彙を用いたプロパティを利用して RDF 形式にしたデータである。

例えば図3のようなクチコミ記事からは、

- foaf:nick ⇒ nick
- foaf:fender ⇒ 男性
- top:age ⇒ 20代
- foaf:name ⇒ 伊豆・三津シーパラダイス
- rev:text ⇒ プラス¥1000で、イルカショーを・・・
- rev:rating ⇒ 4
- dc:date ⇒ 1月

という形でデータをRDFに組み込んでいる。

伊豆・三津シーパラダイスのクチコミ

プラス¥1000で、イルカショーをボートの上から観る「ポートラクション」に参加しました。インストラクターで間近に楽しめてよかったです。記念写真やイルカタッチ等お客様へのサービスがもうちょっとあってよかったかな。

評価	★★★★☆
行った時期	2012年1月21日
投稿者	ゆうさん(男性/20代)

図3: クチコミ記事

2.3 日本語 Wikipedia オントロジーを利用したクラス判定

日本語 Wikipedia から自動的に構築した日本語 Wikipedia オントロジーはその特徴として時事に関わるインスタンスの情報を豊富に含んでいる。日本語 Wikipedia オントロジーのその特性を利用することで、ユーザの入力内容を観光概念に結び付けるのが本モジュールの機能である。

例えば、ユーザが過去に行ったことのある観光地として、「日光東照宮」と入力したとする。「日光東照宮」は静岡県外の観光地であるため静岡観光オントロジーを用いてクラスを調べることはできない。そこでインスタンスについてトピックが豊富である日本語 Wikipedia オントロジーを参照することによって、「日光東照宮」というインスタンスに対して「神社」というクラス(タイプ)が得ることができる。

2.4 日本語 WordNet を利用した概念間距離の測定

ユーザに特定の観光地を推薦するためには、過去に訪れたことのある観光地とRDF化した静岡県のそれぞれのスポットとの関係性を知る必要がある。そこで、日本語に存在する概念を同義語ごとにまとめ、体系化されている日本語 Word Net を用いて、推薦される静岡県内のスポットのクラス概念と過去に訪れた観光地のクラス概念との距離を測定する。

例えば、過去に訪れた観光地が「日光東照宮」、候補に挙げるスポットが「熱海城」とする。日本語 Wikipedia オントロジーを用いて得た「日光東照宮」のクラス「神社」は日本語 WordNet 内では図4のように「宗教建築」のサブクラスであり、宗教建築は「建築物」のサブクラスである。一方、静岡観光オントロジーを用いて得た「熱海城」のクラス「城」は「防御」のサブクラスであり、「防御」は「建築物」のサブクラスであることがわかる。図4のように「神社」クラスから「城」クラスに行くには、4つの辺を通る必要がある。したがってこの二つの概念間の距離は「4」とであると計算することができ、これを「日光東照宮」と「熱海城」の類似性を測定するのに用いる。

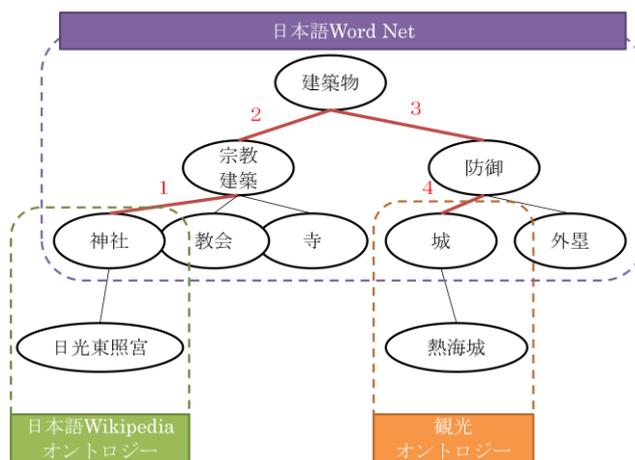


図4: 概念間距離の測定

3. 観光情報推薦システム

今回構築した観光情報推薦システムでは、以下の 2 つのアプローチでユーザへ観光スポットを推薦する。

- ① 口コミ情報に基づくスポット推薦
- ② 日本語 Wikipedia オントロジーと日本語 WordNetを用いたスポット推薦

手順としては、ユーザに自分の情報を入力してもらい、その情報を元にオススメスポットを選定、スポットに関する情報を表示するという流れになっている。推薦に利用する情報と、その属性は次の通りである。

- (1) デモグラフィック属性
 - 性別
 - 年齢
- (2) サイコグラフィック属性
 - 趣味・嗜好
- (3) ビヘイビアル属性
 - 過去に行ってよかった観光スポット
- (4) コンテキスト属性
 - ユーザの位置情報
 - 観光する月
 - 観光の目的
 - 同行者
- (5) (1)~(4)に該当しないが推薦には利用している情報
 - 参考にするクチコミ評価値

① 口コミ情報に基づくスポット推薦

Linked Data として構造化された「性別」、「年齢」、「位置情報」、「観光する月」、「観光目的」、「クチコミ評価値」などの口コミ情報を1つのクエリで SPARQL 検索を行うことによって、ユーザ情報で入力した条件とマッチしたスポットが地図上にアイコンで表示され、スポットの詳細情報を見ることができる。

② 日本語 Wikipedia オントロジーと日本語 WordNet を用いたスポット推薦

キーワードを入力して、そのキーワードをラベル(rdfs:rabel)として持つ、リソースとそのタイプを取得する。各キーワードについては、最初は完全照合を行い、完全照合できなかったキーワードについては部分照合を行う。

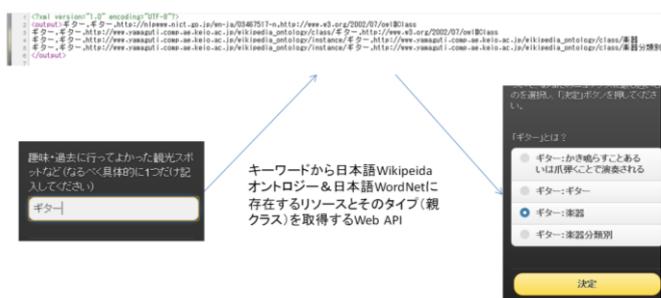


図 5: リソースとタイプの取得

入力クラスと比較対象クラスを入力として、入力クラスと比較対象クラス間の距離を測定し、距離の短い順で上位 5 つを各入力クラスに返す

- ・日本語 Wikipedia と日本語 WordNet の両方のクラスに対応
- ・入力クラスと比較対象クラス間に複数パスが存在する場合には最短パスの距離を採用

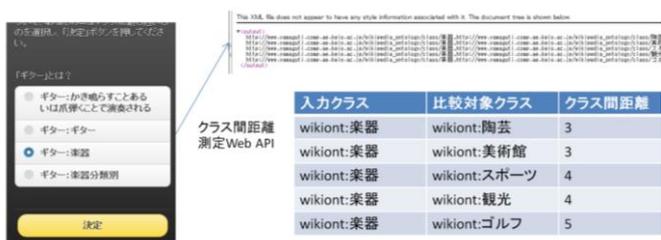


図 6: クラス間距離の測定

ユーザが入力した概念と距離が近いもの上位 5 つに skos:related で関連付けられている観光クラスに属するインスタンスを推薦する観光スポットとして地図上に表示する。

4. 検証と考察

4.1 実証実験

本実験の目的は、作成したアプリケーションが静岡県内で正しく動作するか確認することである。実験は次のような条件で行った。

表2 実証実験におけるユーザ条件

実験場所	静岡県沼津市周辺
エリアの設定	GPSで現在位置を利用
利用目的	見る・遊ぶ
年齢	20代
観光する月	1月
趣味・過去に行って好みのスポット	スポット: 桂離宮 趣味: 楽器

結果としては、スポットの桂離宮から推薦されるスポットとしては、「庭園」クラスに近いクラス概念として「公園」クラスのおすすめスポットが推薦された。これは、ユーザの嗜好に近い推薦結果であった。一方で、趣味の楽器から推薦されるスポットとしては、「楽器」クラスに近いクラス概念として「陶芸」クラスのおすすめスポットが推薦され、趣味とはかけ離れた推薦結果となった。

4.2 既存の観光推薦システムとの比較

従来の観光推薦システムでは、協調フィルタリング (Amazon で実用化されている技術) を利用することにより、ユーザの興味や嗜好といった「サイコグラフィック属性」や行動履歴である「ビヘイビア属性」に基づく情報推薦において秀でている一方、現在

位置や観光する月といった「コンテキスト属性」が推薦のアルゴリズムに反映されていない。

提案システムでは、1で挙げた4つのユーザ属性を考慮した幅広い推薦が行えるという点で既存のシステムに対して優位性があるといえる。一方で、ユーザの嗜好とは必ずしも一致する結果が得られるわけではなく、推薦の精度を上げる必要がある。

4.3 アンケート調査

実証実験に加え、他ユーザにアプリケーションを使ってもらい、その評価を得るためにアンケート調査を行った。アンケート回答者は 11 名、いずれも 20 代男性である。

4.4 結果および考察

実証実験の結果より、本アプリケーションが観光スポットで実用的に動作することが確認された。

一方で、アンケート調査により、機能レベルでのユーザ満足度は決して高いものではなく、改善が必要であることがわかった。特に日本語 Wikipedia オントロジーと日本語 WordNet を用いたスポット推薦機能は、先に述べた概念距離を使ったマッピングだけでなく、オントロジーレベルで人間の感覚に近い概念をつなぐ、といったような工夫が必要であると考えられる。

また、アンケートの中にはユーザ・インターフェースに関する不満や意見も含まれており、今後改善の余地があると思われる。

5. おわりに

本稿では、RDF 形式に変換させた観光情報に関するクチコミ投稿サイトのデータと、オントロジーのクラス-インスタンス関係を利用することで、4種類の属性を持つ情報を組み合わせた観光スポットを推薦するサービスを提案した。そして、提案サービスをスマートフォンで利用するためのアプリケーションを開発し、実際に現地でそのアプリケーションを利用した観光を行うことでアプリケーションの動作性を確認した。観光情報を Linked Data として抽出し、オントロジーと連携した意味処理を行うことにより、従来サービスと比べ、ユーザの嗜好や状況との適合性の高い推薦システムを構築し、推薦システムにおける異なるドメインデータへのリンクやオントロジーとの連携といった Linked Data の有用性を示すことができた。

今後は概念距離を使ったマッピングだけでなく、オントロジーレベルで人間の感覚に近い概念をつなぐ、といったような手法を検討し、ユーザ満足度を上げる必要がある。

参考文献

- [玉川 11] 玉川奨, 森田武史, 山口高平: 日本語 Wikipedia からプロパティを備えたオントロジーの構築, 人工知能学会論文誌 26(4), 504-517, (2011)
- [山本 11] 山本隆三: 観光オントロジーに基づく旅行プラン作成支援, 2010 年度慶應義塾大学大学院理工学研究科開放環境科学専攻修士論文(2011)