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
Shohei NASU
情報メディア学科　2018 年度卒業

1. はじめに

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，自身の興味を拡げることの契機になる．

先行研究では，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が，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．

2. システム構成

アイテムについて，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，抽出したテキストを対象に Mecab を用いて形態素解析を行う．

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れた地名語句集合とする．

地名語句集合について，緯度・経度をジオコーディグを用いて付与する．入力された緯度・経度（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し，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．入力されたアイテムに基づき，他に関連する地域を提示する．

3. 評価

3.1 地域－アイテム推薦

任意の地域を入力したときに，提示されたアイテム集合について定性的に分析する．

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．「甲子園球場」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．

『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コミックが提示された．

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．このことから，沖縄に関連する多くの映画作品やコミックが提示された．

3.2 アイテム－地域推薦

任意のアイテムを入力したとき，提示された地域について定性的に分析する．

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．

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.

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が，関連地域の再現性など問題はいくつか見つかった.

3.3 問題点　アイテム被覆率

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.

3.4 問題点　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.

4. 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を提案した．評価ではアイテム－地名語句推薦と地域－アイテム推薦でそれぞれ定性的に評価し，それによって明らかになった問題点について考察を行った.