Webスクレイピングしたデータを QGIS に適したフォーマットに変換するツールの開発

富山県立大学工学部電子・情報工学科　　指導教員：奥原浩之
1715059　平松楓也

1　はじめに

昨今，人々の生活に多大な影響を及ぼしているコロナウイルスの伝播する様子を視覚的に表現するツールとして QGIS(Quantum Geographic Information System) が用いられている．一般的に QGIS は国や地方が公開しているビックデータを使い地図上に視覚的に表示することに使われている．しかし，QGIS に適した形式のデータを探すことや変換するには時間や手間がかかる問題がある．
ある事柄に対した情報を Web から自動で収集し QGIS に対応したデータに変換するツールはまだ開発されていない．そこで，本研究では，Web スクレイピングしてきたデータを QGIS に適した形式のデータに変換するツールの開発を行う．

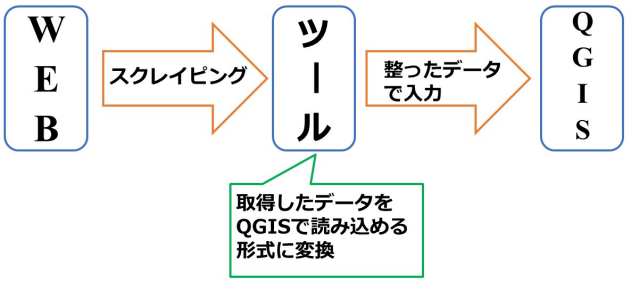

図 1　完成目標

2　QGIS とは

2.1　QGIS の活用例

QGIS はコロナウイルスの感染状況を視覚的に表現したり，ヒグマの出没状況など位置情報が関わるものに広く使われている．

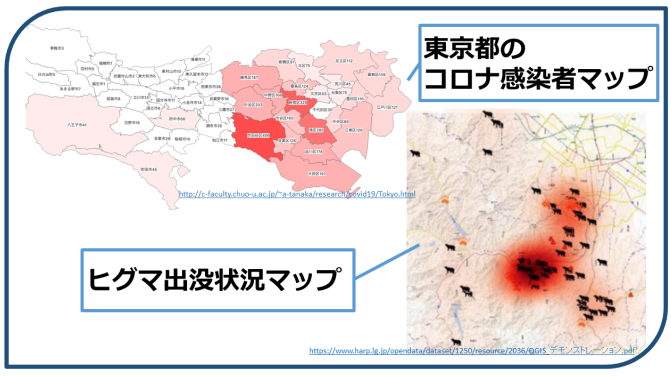

図 2　QGIS の活用例

2.2　データ形式

QGIS は GIS データフォーマットの一つである Shape ファイルというものが使われる．Shape ファイルには道路や建物などの位置や形状，属性データをもつポイント，ライン，ポリゴンで構成されたベクタデータが格納されている．
Shape ファイルは複数のファイルから構成されている．主に，図形の情報を格納する shp ファイル，図形のインデックス情報を格納する shx ファイル，図形の属性情報をテーブルで格納する dbf ファイルの 3 つのファイルで構成される．

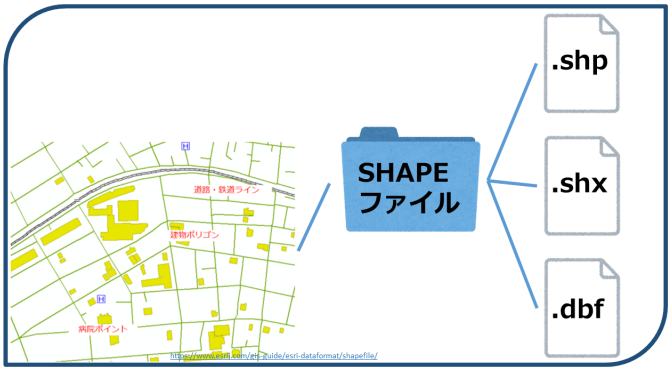

図 3　QGIS の活用例

2.3

3　Web スクレイピング

協調フィルタリングとは，Amazon が開発したレコメンドエンジンで，多くのユーザの嗜好情報を蓄積し，あるユーザと嗜好の類似した他のユーザの情報を用いて自動的に推論を行う方法論である．また，協調フィルタリングには二種類あり，ユーザベース協調フィルタリングとアイテムベース協調フィルタリングがある．

3.1　サイバー空間からのテキストデータの収集

現代社会においてインターネット上の情報量は莫大になっており，今後も増え続けることが予想される．このインターネット上の情報を収集して QGIS で表示することができればより効率的にデータ分析することが可能になると思われる．

今回，キーワードごとに Google 検索から上から何件分かの URL を取得し，その URL からテキストを抽出してそのテキストに対して自然言語処理を行い，必要な単語を取り出す．その単語群から共起ネットワークを作成する．

図 2 にデータ収集に用いるテキストマイニングの説明を示す．

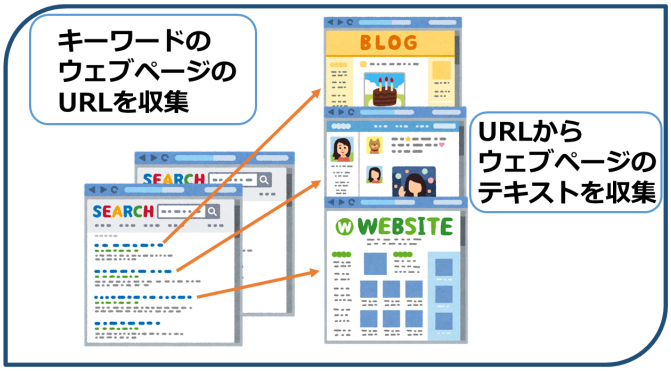

図 4　データ収集の流れ

3.2 テキストマイニング

テキストマイニングとは，大量かつ多量なデータを様々な観点から分析し，役に立つ情報を取り出そうとする技術である．

インターネット上のテキストを用いることで大量のデータを活用することができる．まず，収集した文章に対して HTML タグや JavaScript のコードを取り除くクリーニング処理をする．その次に形態素解析を行う．形態素解析とは文を形態素ごとに分解する技術である [3]．発想支援において必要な名詞や動詞に分解することである．また, 助詞の「は」や「が」助動詞の「です」は不要なので取り除く．そして単語群に対して正規化を行う．正規化することで文字種を統一できる．半角と全角や数字を統一することで同じ単語として扱うことができる．

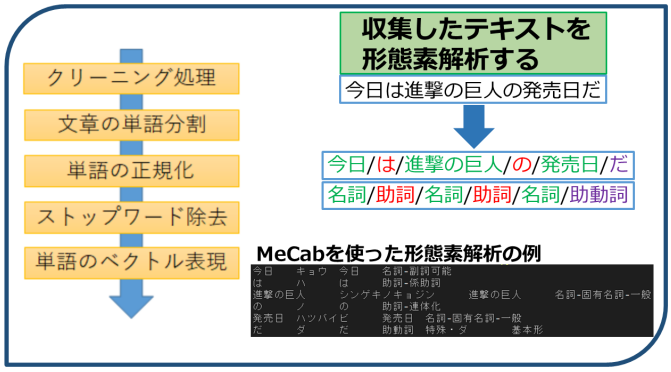

図 5 形態素解析の流れ

4 今回作成するツール

一般に使われる協調フィルタリングは全ユーザのデータを基にフィルタリングを行うのに対し，今回では，ユーザ A が就職を希望している企業に就職したユーザのみでフィルタリングを行い情報推薦を行おうと考えている．

5 進捗状況

・QGIS のお試しとして富山県の色分け人口分布を作成した
・これは国土交通省が公開している shape ファイルのオープンデータの 1 つである行政区域界を少し加工し，富山県が公開している住民基本台帳の csv ファイルのオープンデータを QGIS で結合して作られている
・とりあえず Web スクレイピングしてきたデータを csv ファイルにし，結合させるところまでやって中間を乗り切る予定

6 おわりに

shape ファイルの中身を直接変える方法が見つけられなかったので QGIS 経由で加工している
今後，とりあえずできそうな e-Stat で Web スクレイピングしてみる予定

参考文献

[1] https://www.slideshare.net/takemikami/ss-76817490

[2] https://www.dhbr.net/articles/-/1578?page=3

[3] https://www.digital-knowledge.co.jp/product/edu-ai/edu-ai-merit/

[4] 教学 IR での決定木分析の活用 —初年次の学修成果に影響する入学時の学生特徴の探索を例として— 関西大学高等教育研究 第 8 号 2017 年 3 月