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．

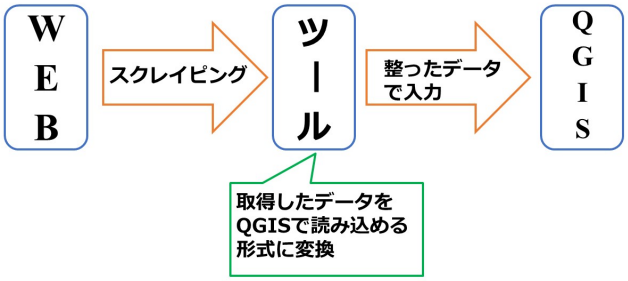

図 1　完成目標

2　QGIS とは

2.1　説明的データ分析

データ分析の中で一番シンプルなものでデータ分析により何か特徴を見つけたり，事実を説明するときに使われる．例えば，どんな人が何を買っているか？ある広告がどれだけ売り上げに貢献しているか？といったことに使われ，その手法は BI，クラスタリング，アソシエーション分析などが挙げられる [1].

2.2　予測的データ分析

未来や欠測値の予測に使われる．例えば，株価やドル円の予測を行ったりすることができる．その手法には，分類・回帰，統計的機械学習，協調フィルタリングのなどが挙げられる [2].

2.3　指示的データ分析

主に最適解を探すことに使われる．例えば，利益を最大化するための，最適な仕入れ量は？などの問題に対し，シミュレーションを行う．また，AI 教育の分野では個別最適化を行い教育の効率化なども取り組まれている．手法としては，最適化やシミュレーション実験などが挙げられる [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　ユーザーベース協調フィルタリング

ユーザベース協調フィルタリングでは「ユーザ A は未評価アイテム I に対して，当該ユーザと似たような嗜好をしている他ユーザと同じような評価をするだろう」という仮定に基づいている．つまりユーザ A と似ている（＝類似度の高い）ユーザの未評価アイテム I への評価点を元にユーザ A の評価点を予測する，というアプローチになる．

3.2　アイテムベース協調フィルタリング

今回用いるアイテムベース協調フィルタリングでは「アイテム同士の類似度とあるユーザ A の過去に評価したアイテムの評価点を用いて未評価アイテム I の評価点を予測する」というアプローチになり，この手法の方がよりオフライン処理しやすく，かつ計算速度という面で優れていることからより使われている [1].

4　今後行うアイテムベース協調フィルタリングについて

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．

拡張子	説明
shp	図形の座標が保存
dbf	属性の情報が保存
shx	shpの図形とdbfの属性の対応関係が保存

図 2　今回の協調フィルタリング

5　進捗状況

協調フィルタリングは制約ボルツマンマシンの応用で進めることになった．制約ボルツマンマシン (Restricted Boltzmann Machine：RBM) とは，入力となる可視層 (visible layer) と隠れ層 (hidden layer) の二層構造のニューラルネットワーク構造を持ち，可視ノード間，隠れノード間には接続がない (条件付き独立性を持つ) ボルツマンマシンである．また，協調フィルタリングに適用する場合は可視変数 r を拡張して考える．可視層を最大評価値の分だけ増やし，選んだ一つに 1 をつけ他は 0 をつける．
RBM の学習とは，与えられた訓練データをもっともうまく説明しそうなパラメータをもつモデルを作成することである．可視層にデータを入れると，隠れ層のフィルタを通して再び可視層に値が返ってくる．その結果を入力と比較し最適な出力が得られる方向にパラメータを調整する．しかし，このパラメータ更新の計算をする際に組み合わせ爆発の問題があるため，何らかのサンプリングを行いその平均で期待値を近似する必要がある．近似には CD 法を用いる．CD 法とは，ギブスサンプリングより処理を早くしたもので，サンプリングには MCMC(マルコフ連鎖モンテカルロ法) を用いる．

6　おわりに

中国の人が python で制約ボルツマンマシンを使った協調フィルタリングのコードを見つけた．現在コードを読んでる最中で，購入した本も参考にすれば python のプログラムの作成はできると思う．

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