機械学習による学習データの収集

プロジェクトマネジメントコースソフトウェア開発管理グループ　矢吹研究室　1442031　小山隆太郎

1. 序論

私たちは自宅に居ながら, あるいは手元の端末で世界規模の地図や気象情報すら閲覧できるようになった．インターネットが普及して 20 年が経った今, パソコンやスマートフォンなどを利用してできる様々なサービスは私たちの生活に欠かせないものとなっている．そのほとんどのサービスを利用するのに私たちは「検索する」機能を利用する．私たちユーザーは自分が解決したい情報を探そうと検索窓にワードを入力していく．すると, 検索エンジンは世界中の膨大なウェブページの中から, 入力された言葉を含み, 問題解決に役立ちそうなウェブページを瞬時に選び出して検索結果として返してくれる.[1]　しかし多くのウェブページの中から検索することは, 結果として意図のないウェブページを含んでしまう．ユーザーの意図に合ったウェブページを見つけ出すためには, ユーザーと検索したウェブページの関係だけではなく, ユーザーの性別や年齢・趣味やこれまでどんな検索を行ってきたのかなどの多くの要因を組み合わせて, ユーザーに合ったウェブページを見つけ出さなければならない.[2]　これらの要因をどのように組み合わせれば, ユーザーの意図に合ったウェブページを見つけることができるのだろうか．ここには機械学習と呼ばれる技術が活用されている．機械学習とは人工知能の一分野であり, 入力した言葉を機械自身に学習させるという手法である．ユーザーの要因をどのように組み合わせればよいのかはわからないが, 検索する言葉とウェブページが与えられたとき, どのウェブページが好ましいか判断することができるようになる．この判断結果をもとに機械は最適な要因の組み合わせを学習する.[3]　機械学習はこれらの学習データを大量に読み込み, そこからどのように要因を組み合わせたら, 学習データの判断を真似することができるのかを判断する．そしてまた検索をするときには機械学習によって得られた要因の組み合わせ方に従い, 検索するウェブページの検索結果を決定する．

2. 目的

私の課題研究の企画は, 学習データをいくつか試しに自分で集め, 特定のユーザーの学習データを得られるようにする．得られた結果からユーザーに最適な検索結果を表示できるようにし, この企画の実行を通して実世界での活用法を模索していく．

3. 手法

1. 例として Amazon を挙げ,jQuery を使えるようにする．

2.html から注文データを抽出する．

3. 注文履歴のある URL を繰り返し取得する再帰処理を行う．

参考文献

[1] 高野明彦. 検索の新地平. 角川学芸出版, 第 8 版, 2015.

[2] 東浩紀. 弱いつながり　検索ワードを探す旅. 幻冬舎, 第 01 版, 2014.

[3] 石井康之．ビックデータ処理技術を用いた wikipedia マイニング．千葉工業大学卒業論文 2015, 2015. 117.