ウェブスクレイピング AI の衝撃

～ここまで来た！ウェブ情報取得の世界～

ライフデザイン研究部 主席研究員 柏村 祐

1.ウェブスクレイピングとは

ビジネスにおいては、さまざまな局面で業界の最新動向などの情報をウェブで調査する機会がある。それらの情報を自社の資料に活用するには、ウェブページから 1 つ 1 つ情報を転記・加工する作業が必要となるが、その作業を担うソフトウェア技術にウェブスクレイピングがある。

英語の scraping（削ること、こすること）に由来するウェブスクレイピングは、商品価格の比較、口コミの分析、株価の監視などさまざまな用途に活用されている。ウェブスクレイピングは、民間企業のみならず行政も活用している。例えば、総務省統計局では、CPI（消費者物価指数）の基準改定にあたり、ウェブ上に公開されているネット価格の収集拡大に向け、データ検証を重ねてきた。その結果を踏まえ、2020 年の基準改定においては、ネット購入割合が高い旅行サービスについて、ウェブスクレイピングを活用してネット価格を網羅的に収集している（注 1）。

現在、ウェブスクレイピングを行うには、専門的なプログラムを作成し、ウェブページの情報を取得する必要がある。だが、プログラミング人材が乏しい組織においては、ウェブ上の情報を 1 つ 1 つ手作業で転記したり、コピー＆ペーストする生産性の低い手法が用いられている。このような手作業による作業時間の短縮や転記ミスの防止につながる仕組みとしてウェブスクレイピング AI が登場している。

本稿では、そのウェブスクレイピング AI について概観し、その可能性について解説する。

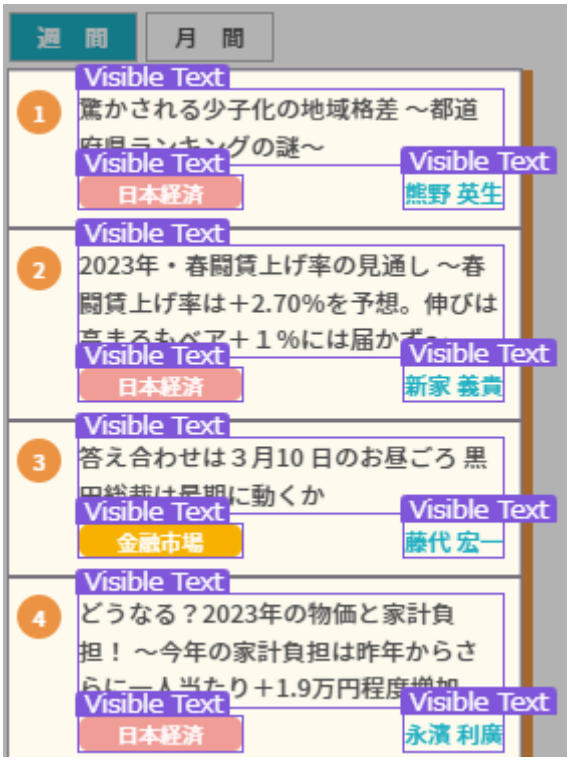
2.ウェブスクレイピング AI とは

ウェブスクレイピングは、ウェブサイトから特定のデータを抽出する仕組みで、製品価格調査、EC サイトからの商品・価格情報の抽出、自社のビジネス情報の収集など多様な分野で活用されている。ウェブスクレイピングを行うには、スクレイピング作業に必要なプログラミング言語でプログラムを作成する必要があるが、ウェブスクレイピング AI を用いれば、ワードやエクセルのようなソフトウェアを操作する感覚で、ウェブ上に掲載されている情報を編集可能なデータとして出力できる。

そこで、具体的なケースとして当社のホームページに公開されている「レポートランキング情報」と「新着レポート情報」を題材とし、実際にウェブスクレイピング AI を動作させ、ウェブサイトから特定のデータの抽出を試みた。

まず、当社ホームページの「レポートランキング情報」を題材として、ウェブスクレイピング AI の性能を検証してみた。ウェブスクレイピング AI を使えば、プログラミングのコードは必要がなく、マウス操作とテキスト入力作業だけでスクレイピングを実行できる。ウェブスクレイピング AI を起動させ、スクレイピングの対象となる当社のホームページの URL を指定する。その後、今回スクレイプングしたいウェブ上の場所となる週間ランキングの欄を範囲選択する。そして、週間レポートランキングに表示されているタイトル、カテゴリ、発信者にそれぞれカーソルをフォーカスさせると、ウェブスクレイピング AI は、それぞれの対象を自動認識してくれる（図表 1）。

図表 1　AI がスクレイピングしたいランキング情報の項目を自動認識する様子



資料：browse.aiHP「https://www.browse.ai/」より筆者作成

それぞれ対象として認識された情報には、テキストを入力できるポップアップが表示されるため（図表 1 の Visible Text の部分）、付与したいタグ名を記入すればよい。今回は、ウェブスクレイピング AI が認識した情報に対するタグ名として、title（タイトル）、category（カテゴリ）、writer（発信者）を入力している。その後、ウェブスクレイピング AI の機能である完了ボタンを押せば、AI は瞬時にワードやエクセルのようなソフトウェア上で並べ替えたり、加工できる編集可能データを生成してくれる（図表 2）。

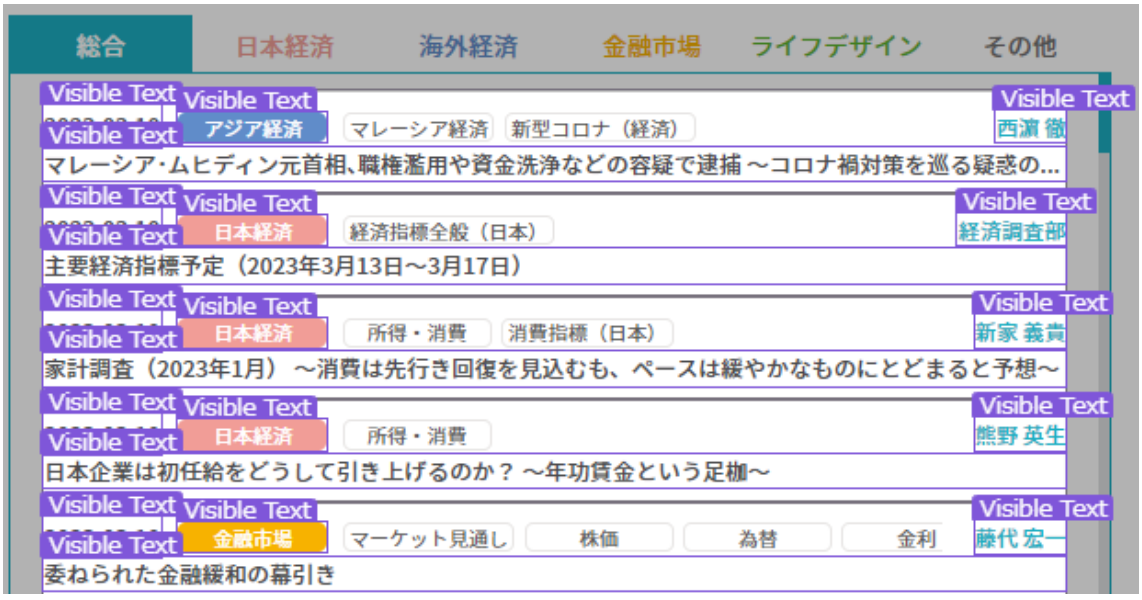
図表 2　AI がランキング情報を編集可能データにする様子

Position	title	category	writer
1	驚かされる少子化の地域格差 ～都道府県ランキングの謎～	日本経済	熊野 英生
2	2023年・春闘賃上げ率の見通し ～春闘賃上げ率は＋2.70%を予想。伸びは高まるもベア＋1%には届かず～	日本経済	新家 義貴
3	答え合わせは３月10 日のお昼ごろ 黒田総裁は最期に動くか	金融市場	藤代 宏一
4	どうなる？2023年の物価と家計負担！ ～今年の家計負担は昨年からさらに一人当たり＋1.9万円程度増加の可能性～	日本経済	永濱 利廣

資料：browse.aiHP「https://www.browse.ai/」より筆者作成

次に、日々発信される「新着レポート情報」についてもウェブスクレイピング AI を動作させ、その性能を検証してみた。当社ホームページに掲載される新着レポートのウェブページ上でウェブスクレイピング AI を起動させる。その後、今回スクレイピングしたいウェブ上の新着レポート欄に表示されている日付、カテゴリ、筆者、タイトルにそれぞれカーソルをフォーカスさせると、ウェブスクレイピング AI は、それぞれの対象を自動認識してくれる（図表 3）。

図表 3　AI がスクレイピングしたい新着レポートの項目を自動認識する様子



資料：browse.aiHP「https://www.browse.ai/」より筆者作成

データ生成の対象として認識された情報には、テキストを入力できるポップアップ

が表示されるため（図表 3 の Visible Text）、付与したいタグ名を記入すればよい。今回の新着レポートに関してウェブスクレイピング AI が認識した情報に対して、date（日付）、category（カテゴリ）、writer（発信者）、title（タイトル）を入力している。その後、ウェブスクレイピング AI の生成機能を発動させれば、AI は瞬時にワードやエクセルのようなソフトウェア上で並べ替えたり、加工できる編集可能データを生成してくれる（図表 4）。

図表 4　AI が新着レポート情報を編集可能データにする様子

date	category	writer	title
2023.03.10	アジア経済	西濵 徹	マレーシア･ムヒディン元首相、職権濫用や資金洗浄などの容疑で逮捕 ～コロナ禍対策を巡る疑惑の一方で政局争いが影響した可能性も、政治の成熟化は望めないか～
2023.03.10	日本経済	経済調査部	主要経済指標予定（2023年3月13日～3月17日）
2023.03.10	日本経済	新家 義貴	家計調査（2023年1月） ～消費は先行き回復を見込むも、ペースは緩やかなものにとどまると予想～
2023.03.10	日本経済	熊野 英生	日本企業は初任給をどうして引き上げるのか？ ～年功賃金という足枷～
2023.03.10	金融市場	藤代 宏一	委ねられた金融緩和の幕引き

資料：browse.aiHP「https://www.browse.ai/」より筆者作成

3.ウェブスクレイピング AI の可能性

以上のように、ウェブスクレイピング AI は、誰もが行える簡単な作業を加えることで、ウェブ上から抽出した情報を編集可能なデータとして生成してくれる。この AI は、プログラミングスキルをもつ一部の人にしかできないと思われているウェブ情報のデータ化について、特段そのスキルがない人でも容易に行える世界を実現している。

現在、ウェブスクレイピングが必要な場合は、社外の専門家に対価を払ってデータ生成を発注したり、プログラミングスキルをもつ社内の人材に依頼する必要がある。今後、ウェブから情報を取得する必要が生じた際に、このウェブスクレイピング AI を活用すれば、個人や組織の生産性を飛躍的に向上させるだろう。ただ活用の際には、情報取得元であるウェブサイトの利用規約、著作権や個人情報保護などの法的な問題に十分配慮する必要がある。

なお、現時点ではエッジやクロームといった検索エンジンにウェブスクレイピング AI の機能は搭載されていない。そのため、その機能を活用するにはプラグインをインストールする必要があるが、今後検索エンジンの基本機能として搭載されることも考えられる。

ビジネスで必須ともいえるウェブからの情報収集作業において、ウェブスクレイピング AI は、生産性を向上させる AI の 1 つとして今後さらに進化していくであろう。

【注釈】

1）総務省統計局 HP より
https://www.stat.go.jp/info/today/148.html