未来へつなぐ新聞ビッグデータ

Newspaper Big Data for the Future

前田祥 Maeda Akira

1. 背景

本プロジェクトは、北海道新聞社から提供を受けた新聞ビッグデータを利用して新聞のテキストを可視化し情報を提供する「View Picks」というWeb アプリケーションを開発した。現在のマスコミュニケーションの手段として新聞以外に SNS やインターネットが挙げられる。しかし、これらのマスコミュニケーションの手段は、利用するユーザーに最適化された情報を多く提供するため、享受する知識が隔たってしまうという問 題点がある。しかし、新聞やテレビ、ラジオといったマスコミュニケーションの手段では、街 角の事件・事故やゴシップ、風俗、風潮、広告、連載漫画、小説など幅広いコンテンツの知識 を享受することができる。その中でも新聞は、ジャンルを問わず自ずと求めていない知識を享受することが可能である。この新聞の特性に着目し、新聞に利用された言葉とインタラクティブに触れる体験を行うことが可能なWeb アプリケーションの利用を通し、可視化された内容を読み取り気に留まった言葉を検索し新聞に記載された出来事へと繋げ新聞に対する興味関心を向上させることが狙いである。

2. 課題解決のプロセスとその結果

新聞について調査、課題を設定、課題を解決するプロダクトを作成という過程を経た。

2.1 新聞の現状調査

2.1.1 新聞の社会的位置

4 大マスメディアのうちの一つである新聞は、古くから様々な人に利用されてきており、幅広い 分野に関する情報を受け取ることができる情報源として、我々の生活に欠かせないものであった。日本新聞協会[1]によると、2000 年における一般紙とスポーツ紙を合 わせた発行部数が約 5400 万部であったのに対し、2021 年は約 3300 万部と、およそ 2100 万部減少していることが分かった。原因として SNS の普及が挙げられる。

2.1.2 新聞の利点

新聞の利点として、次の 2 つを挙げる。一つ目の利点は、一覧性が高いことである。二つ目は、信頼性が高いことである。図 2.12 で示す総務省の令和 3 年度版情報通信白書[2]では、「信頼できる」と回答した人の割合は、新聞の 61.2%が最も多い。一方、本プロジェクトのメンバーのような若い世代が日常的に利用する SNS[3]や、その他インターネットを利用したメディアの信頼性については以下のような結果になった。SNS（15.3%）、動画投稿・共有サイト（14.4%）、ブログなどその他サイト（10.3%）、掲示板やフォーラム（7.3%）となっており、上記で挙げた新聞・テレビ・ラジオよりも信頼性が低いことが分かった。

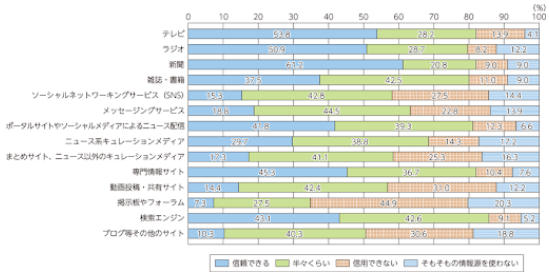

図 2.1.2　各メディアに対する信頼性（出典:令和 3 年度版情報通信白書）

2.1.3 新聞の課題

新聞の課題として、次の 2 つを挙げる。一つ目は、ユーザーが見たい情報だけを効率よく得ることができないということである。二つ目は、情報拡散能力が低いということである。テレビや SNS では、何か大きな事故が起きると、‘速報’という形ですぐに情報が拡散されるため、時間をかけることなく最新のニュースを取得することができる。

2.2 新聞の解決すべき点

ユーザーが見たい情報だけを効率よく得ることができないということ、そして、情報の 速報性に乏しいという新聞に対する 2 つの課題を挙げた。新聞は、大量の情報があるゆえ に自分が求めている情報だけを効率よく得ることができない一方、そのメリットである一覧性を生かすことで、興味のなかった分野の記事や自分が知らなかった情報を新たに得ることができるという強みを改善策に活かすことが可能である。そこで今回は、この改善策に対して、より具体的な取り組みを考えることで新聞の新しい使い方を提案する。具体的な取り組みとしては、一覧性が高い新聞のテキストデータを可視化した新聞の情報を新聞に普段触れることのないユーザーに対して提供するということである。本プロジェクトでの成果物は新聞記事内に含まれるテキストデータや、広告などといった大量の情報が一目で分りやすく可視化されていることにより、ユーザーがあまり興味のない情報でも受け取りやすいといった利点がある。また、ユーザーにただ新聞記事を読んでもらうのではなく、様々な方法での可視化を通すことで、新聞に眠っている情報の価値や面白さを理解してもらうことができる。

2.3 アイデア出し

成果物案決定のために図 2.3 のようにビッグデータ、新聞の活用方法についてブレインストーミングを利用し製作するプロダクトのアイデア出しに努めた。2 グループに分かれ、それぞれアイデア出しを務めた。最終的に 2 グループの案を統合し、ビッグデータを可視化するプロダクトを成果物案とした。その後は、データを可視化する際の
表示方法や作成する意味などをグループ全体で話し合い成果物の最終案を決定した。

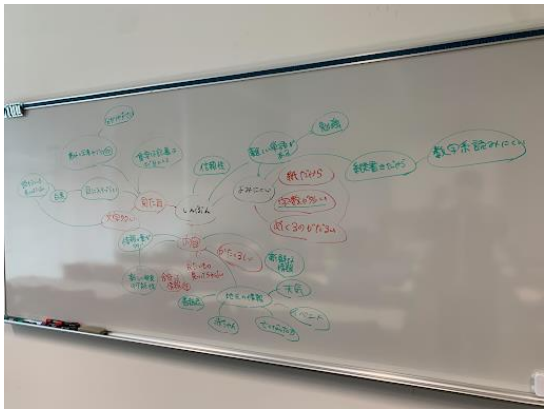

図 2.3　ブレインストーミングの様子

2.4 成果物の開発

2.4.1 成果物の開発する際の役割り分担

成果物の案が決定した後に、必要となる技術やプラットフォームを洗いだした。その結果、デザイナー、Web フロントエンドエンジニア、バックエンドエンジニア、データエンジニアリング、画像エンジニアの 5 役職が必要であることが分かった。後期での活動で役割りを遂行するため、夏季休暇を利用しインターンシップや自己学習で技術の習得に励んだ。

2.4.2 開発手法

開発手法では短期間で効率的に開発を行うために、図2.4.2に示すようにスクラム開発を採用した。GitHub でスプリントを管理し、スクラムリーダーを筆頭に開発を推し進めた。その結果、各役割りでの連携、プロトタイプの開発を効率的に行うことができ開発が滞る問題を解消することができた。

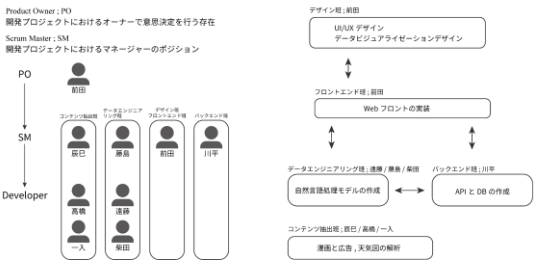

図 2.4.2　スクラム開発の組織図

2.4.3 デザイナーの役割り

デザイナーは Web アプリケーションの UI、UX、データビジュアライゼーションデザインを担当した。UI には OOUI（オブジェクト指向ユーザーインターフェース）を採用し、わかりやすい情報設計を目指した。

2.4.4 Web フロントエンドエンジニアの役割り

Web フロントエンジニアはビッグデータを可視化するために、Web アプリケーション上にグラフィックを可視化するシステムの開発を担当した。D3.js という JavaScript の可視化ライブラリを選択し、可視化に必要な API をバックエンド担当、データエンジニアリング担当と連携し認識の齟齬が生じないように努めた。

2.4.5 バックエンドエンジニアの役割り

バックエンドエンジニアはバックエンドシステ

ムの構築と API の開発を担当した。システムは Linux マシン上の Docker を用いて本番環境を構築した。Web アプリケーション業界には Linux、Apache、MySQL、PHP の頭文字をとった LAMP という王道的な構成が存在する。しかし、開発期間が短いこと、機能拡張が容易である必要があることなどといった理由から LAMP は無視することとした。その結果、Web サーバー兼リバースプロキシには Nginx、アプリケーションサーバーには Nginx Unit、リレーショナルデータベースには MySQL を、そして API を実装するための言語には Python を用いた。実際に実装した API から取得する json データの形式は図 2.4.5 の通りである。またビッグデータを高速に処理するためにデータの操作を全てデータベースで完結させた。また API の応答速度を向上させるために、オンメモリキャッシュを Redis で導入し API のレスポンス速度を 10 秒程度短くした。


```
[
    {"date":"1994-07-01","word":"自民党","count":80},
    {"date":"1994-07-01","word":"社会党","count":79},
    {"date":"1994-07-01","word":"委員長","count":42},
    {"date":"1994-07-01","word":"課長補佐","count":34},
    {"date":"1994-07-01","word":"事務所","count":30}
]
```


図 2.4.5　API から取得する json データの形式

2.4.6 データエンジニアの役割り

データエンジニアは新聞のビッグデータが含まれた CSV ファイル形式のデータを自然言語処理で解析しアプリケーションの API で利用する json データの生成を担当した。自然言語処理を行うための開発環境として、Google Colaboratory を使用した。データの整形・高速化を行うために numpy や glob モジュールを用いて開発を進めた。numpy は記事内に出現した単語間のコサイン類似度を求める際に活用し、glob モジュールは、引数に指定されたパターンにマッチする記事のファイルパス名を取得することに活用できた。さらに、関連語を出力する際に、比較的データサイズが小さい日本語の学習済み Word2vec モデルを使用した。これらの手法を用いた結果、計算の高速化とデータサイズの縮小化に成功した。

2.4.7 画像エンジニアの役割り

画像エンジニアは、YOLO（You Only Look Once）という高速な物体検出アルゴリズムを用いて新聞の 4 コマ漫画などのコンテンツ検出を担当した。4 コマ漫画の学習データセットを自作し、機械学習を行った。図 2.4.7.1 の学習結果が得られた。YOLO での学習を進めていき、最終的には一定以上の精度で新聞の画像データから、4 コマ漫画と天気コーナーの検出に成功した。図 2.4.7.2 はそれらを検出することに成功した結果である。左上に 4 コマ漫画、左下に天気コーナーを検出することに成功したことがわかる。

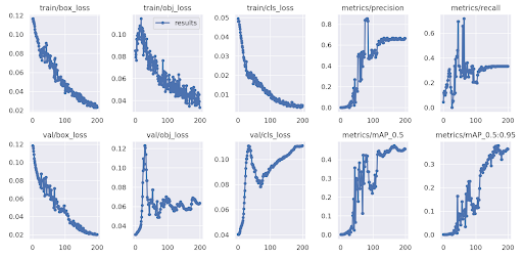

図 2.4.7.1　学習結果のグラフ

図 2.4.7.2　検出結果

2.5 最終成果物

新聞のテキストデータを可視化し知的好奇心を刺激する”View Picks”という Web アプリケーションを開発した。本 Web アプリケーションでは、日付の区間を選択すると、選択した日付の区間に含まれる新聞の内容が可視化される。そのため、情報の強弱や関係性をインタラクティブに体験し知的好奇心を刺激する体験を提供することができる。本プロジェクトの開発目的であった「知らなかった情報を知ることができる機会を作り、知的好奇心を刺激できるようなものを作る。」を達成するた

めに作成したものであり、新聞に眠る情報の価値を引き出し、 マスコミュニケーションの手段としての新聞ではなく教育への活用や知的好奇心を刺激する方向へと導くなどの目的に通じる Web アプリケーションである。 また、この Web アプリケーションのシステム構成は図 2.5.1 の通りとなっている。図 2.5.2 はアプリケーションのホーム画面、図 2.5.3 はアプリケーションの可視化画面である。

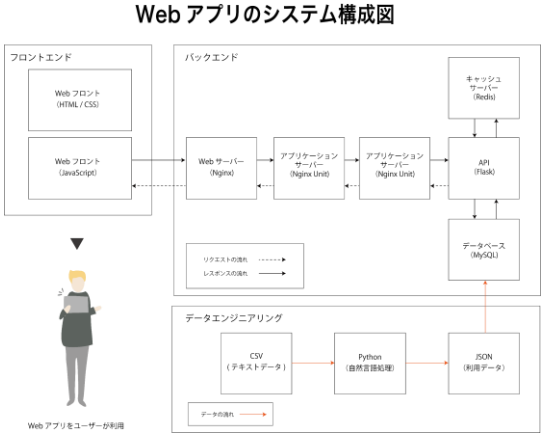

図 2.5.1　Web アプリケーションのシステム構成図

図 2.5.2　Web アプリケーションのホーム画面

図 2.5.3　Web アプリケーションのホーム画面

3. まとめと今後の課題

本プロジェクトは、アプリケーションを作成し、それが新聞に眠る情報の価値を引き出すものであり、マスコミュニケーションの手段としての新聞ではなく教育への活用や知的好奇心を刺激する方向へと導き、新聞の新しい使い方を提案する可能性を示唆することを目的としていた。そして私たちはその目的を達成できたと考えている。View Picks を使い、情報を様々な方法で可視化したことにより、普通に新聞の紙面を読むことで得られる情報とはまた違った情報を手に入れることができたためである。今後の課題としては、可視化の種類の増加、教育での利用を想定したアプリケーションの改善などが挙げられる。例えば、ワードクラウドの漢字に自動的にフリガナを適用させることが考えられる。

謝辞

本プロジェクトでは北海道新聞の過去 33 年分の新聞記事データを活用した。プロジェクトの構想やシステム試作において北海道新聞函館支社長の三浦辰治氏に多大なるご協力を頂戴したので感謝申し上げる。

参考文献

[1] 日本新聞協会（2022） 新聞の発行部数と世帯数の推移. https://www.pressnet.or.jp/data/circulation/circulation01.php (2023/01/04 アクセス)

[2] 総務省（2021） 令和 3 年度版　情報通信白書, pp.218-219, https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r03/pdf/01honpen.pdf

[3] 株式会社 NTT ドコモ モバイル社会研究所（2022）モバイル社会白書　2022 年版, p.87, https://www.moba-ken.jp/whitepaper/wp22/pdf/wp22_all.pdf