WWW 横断型ヒューマンコンピュテーション

白井 良成 * 岸野 泰恵 * 柳沢 豊 * 水谷 伸 * 須山 敬之 *

概要. WWW 上に存在するコンテンツは膨大かつ多種多様であり，機械学習のための学習データとして利用可能な可能性を秘めている．しかし，WWW 上から学習データとして利用可能なコンテンツを探し出し，機械学習用に正解ラベルを付与する作業は，多くのリソースを必要とする．そこで，本稿では，WWW の利用者をヒューマンコンピュテーションのワーカと見做し，任意の Web ページに対するラベリングを促進する概念，WWW 横断型ヒューマンコンピュテーションを提案する．既存の Web ブラウザに対してラベリング機能を組み込むことで 3 種類のシステムを実装し概念の実現性を確認した．

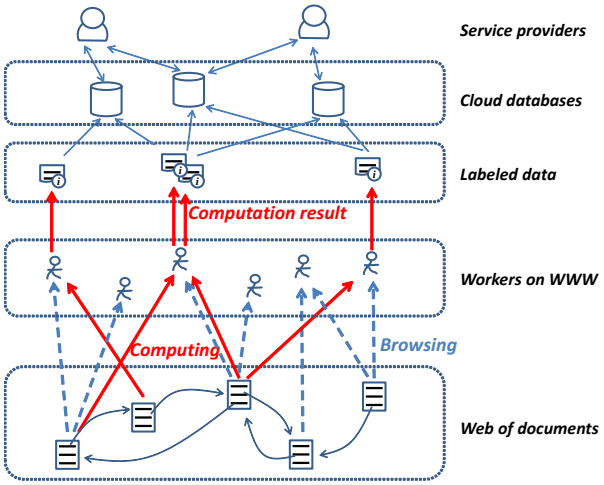


図 1. WWW 横断型ヒューマンコンピュテーション

1 WWW 横断型 HC

社会実装の進む深層学習を効果的に利用するためには，大量の学習データ（正解ラベル付きデータ）を準備する必要がある．通常学習データは，人手で準備する必要があり，深層学習等の機械学習を利用したサービスを実現する上での大きなボトルネックとなる．この課題を解消するため，様々な手法が提案されている [3, 1]．また，不特定多数のワーカに依頼可能なクラウドソーシングサービスも近年では利用可能である．しかし，いずれの方法を用いても十分な学習データの作成には，依然として無視できないコスト (人的／金銭的リソース) が必要となる．

我々は，WWW が機械学習の学習データを作成するフィールドとして優れていると考えている．WWW にはデジタル化された膨大かつ多種多様なデータが存在し，また日々追加更新されている．どのような学習データを必要としているかによるものの，目的の学習データの作成に利用できる素材はかなりの確率で WWW 上に一定量存在すると思われる．また，現代人の多くは，日常生活においてかなりの時間 WWW 上での活動 (Web ページの検索，閲覧，作成など) に従事している．WWW 上における活動時間の一部を学習データ作成作業に割いてもらえれば，たとえそれが WWW 利用者のごくわずかであったとしても，かなりの学習データを作成できよう．

このような考えに基づき，本論では WWW を利用することで学習データの効率作成を実現する概念，WWW 横断型ヒューマンコンピュテーション (WWW 横断型 HC) を提案する．WWW 横断型 HC の概念モデルを図 1 に示す．WWW 横断型 HC では，WWW 上で活動している人々を HC の潜在的なワーカと見做す．WWW 上で活動している人の一部が WWW 上のデータにラベルを付与していくことで，大量のラベル付きデータが作成される．作成されたラベル付きデータは，サービス提供者が AI を用いたサービスを実装する際の機械学習用学習データ等に利用することを想定する．

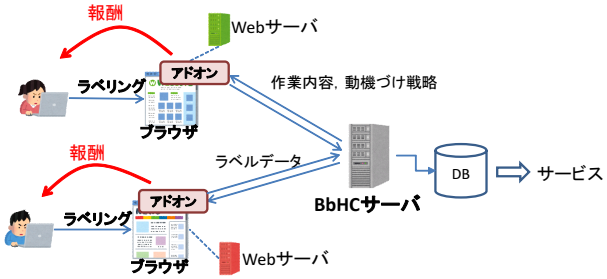
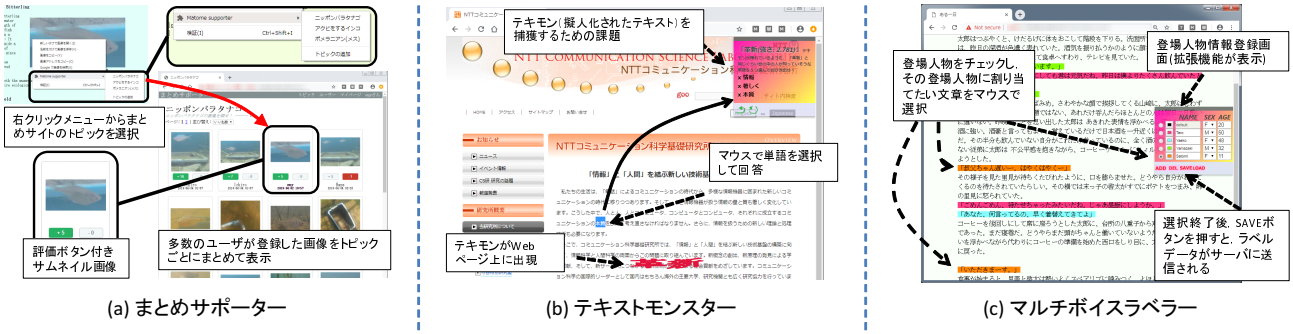


図 2. Browser-based Human Computation

2 Browser-based Human Computation

WWW 横断型 HC の概念を実現するために，我々は，既存の Web ブラウザを HC の作業ツールとすることとした．本アプローチを BbHC(Browser-based

* 日本電信電話 (株) コミュニケーション科学基礎研究所



(a) まとめサポーター　(b) テキストモンスター　(c) マルチボイスラベラー

図 3. BbHC に基づくシステム群

Human Computation) と呼ぶ．Chrome や Firefox 等の Web ブラウザは，多くの人が日常的に Web ページの検索や閲覧に利用しており，WWW 上で活動する人をワーカと見做す WWW 横断型 HC の実現に適していると考えた．Web ブラウザを HC 作業ツールとすることで，ユーザは Web 閲覧中に HC 作業に適したデータを見つけた際に，素早くその作業に従事することができるだろう．

既存の Web ブラウザは現在のところ HC 作業の機能は有していないため，BbHC では，ブラウザの拡張機能 (アドオン) を利用して Web ブラウザを HC 作業ツール化する (図 2)．アドオンは作業インタフェースを提供する．作業結果は，BbHC サーバを介してデータベースにラベル付きデータとして保存される．動機づけと品質管理は他の HC システムと同様 BbHC においても重要である．アドオンはサーバと連携しながら Web ブラウザ利用者に魅力的な報酬を提供し，HC への従事を促すべきである．また，作業結果の品質を評価し，学習データとして必要十分な品質を確保する必要がある．

BbHC の利点は，学習データの作成に必要な素材を準備せずに，ラベルの付与作業を人々に依頼できる点だ．例えば，紙媒体の印字画像へのラベル付与を実現した reCAPTHA[5] や銀河の画像を提示してユーザに形態を分類してもらう GalaxyZoo[4] では，印字のスキャン画像や銀河の画像を依頼者が準備する必要がある．一方，BbHC ではユーザが閲覧している Web ページが，ラベリング素材そのものであり，ラベリング素材を準備する必要が無い．BbHC で依頼するラベリング作業には，ラベリング素材を WWW 上から捜索する作業が暗黙裡に含まれる．

3　BbHC に基づくシステムの実装

BbHC に基づき 3 種類のシステムを実装することで，WWW 横断型 HC の実現性を確認した (図 3)．まとめサポーターは，「画像のまとめサイト」を共同で簡単に作成できるシステムである．ユーザによって収集されたラベル付き画像群は，画像識別器を作成するための訓練データとして利用できる．テキストモンスターは，擬人化されたテキストを用いて Web サイトを奪い合うゲームである．サブゲームの結果を通して，単語親密度データベース [2] を更新することができる．マルチボイスラベラーは，WWW 上の小説等を複数のボイスで適切に読み上げるためのアノテーションを付与するツールである．アノテーションが付与された Web ページは，話者情報付き言語資源として，談話構造解析の研究などに活用できる．各システムのラベリング機能は Chrome 拡張機能によって実現し，BbHC サーバは node.js+PostgreSQL によって実装した．

4　おわりに

本稿では WWW 横断型 HC の概念，それを実現する BbHC アプローチを提案した．また，BbHC に基づいて構築した 3 種類のシステムを紹介し，WWW をフィールドとして多様な HC システムが構築できることを示した．

参考文献

[1] D. Acuna, et al. Efficient Interactive Annotation of Segmentation Datasets With Polygon-RNN++. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] S. Amano and T. Kondo. Estimation of mental lexicon size with word familiarity database. In *The 5th International Conference on Spoken Language Processing*, 1998.

[3] M. Andriluka, et al. Fluid Annotation: A Human-Machine Collaboration Interface for Full Image Annotation. In *Proc. MM2018*, pp. 1957–1966, 2018.

[4] C. J. Lintott, et al. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008.

[5] von Ahn, et al. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895):1465–1468, 2008.