アンケート調査と行動履歴を統合させた高精度な ユーザ属性推定プラットフォームの提案

清水 伸幸[†]　　坪内 孝太[†]

本稿ではクラウドソーシングを用いたアンケート結果と、検索ログを組み合わせた計算社会科学調査のプラットフォームを提案する。　利用例としては、興味推定が挙げられる。　Yahoo! JAPAN が持つマイクロタスク型のクラウドソーシングサービスを利用して興味推定のアンケート調査を行い、ユーザ ID を検索ログと結合することで、アンケート結果を持たない検索ログのみ入手可能なユーザについても興味を推定することが可能となる。　本プラットフォームの適用範囲は興味推定にかぎらず、デモグラフィックデータの推定などにも応用可能であり、社会科学分野での課題解決へ向けた活用が有望視されている。

キーワード：クラウドソーシング, アンケート調査, ユーザ属性推定

1 はじめに

近年、人々のオンライン上での活動データを情報技術によって解析し、オフライン（実世界）の行動や社会現象の理解に役立てよう、という動きが広まっている。　それに伴い、従来別々のものと考えられていた複数の学術分野にまたがる「計算社会科学」(Computational Social Science)という学際分野が生まれ、新たに注目を集めている。　この計算社会科学の分野で特に有望なアプローチとして、オンライン上での人々の自発的な情報収集行動や、コミュニケーションのデジタルトレース（ビッグデータ）を使い、伝統的な社会科学で行われてきたアンケート調査を補完する手法が挙げられる。　[Blumenstock et al., 2015]の既存研究では、携帯電話のユーザに対してアンケートをとり、結果から得られた対象ユーザと携帯電話のメタデータを持つ関係を機械学習で見つけ出すことで、携帯メタデータのみしか得られないユーザに対しても貧困状態などのデモグラフィックデータを推測することが可能となった。

これまで、計算広告(Computational Advertisement)や推薦システムの分野では、「クリックした」「検索した」「商品を購入した」「資料請求をした」などのビッグデータと呼ばれるサービスの各種ログのみを用いて、ユーザの興味を推定する試みが広く行われてきた。　だが、これらのログはアンケートと比べ、量は多いものの得たい対象データそのものではなく、対象データに変換するために加工が必要となる上、ノイズの問題がある。　一方、通常のアンケート調査では、対象とするデータが直接得られるものの、調査コストが大きく、調査範囲が限られるという問題がある。　したがって、ビッグデータとアンケート調査は互いに補完し合う関係にある。　だが、ビッグデータとアンケート調査を組み合わせるにあたっては、サービスのログとアンケートの ID が紐付かないという課題がある。　そのため、実際にアンケートを行うのではなく、例えば年齢や性別といったサービスへの登録情報の回答結果から属性毎に商品に

[†]ヤフー株式会社 Yahoo! JAPAN 研究所

興味をもつ確率を推測し、当てはめる解法 [Cheng et al., 2012]が現実的であった。　しかしながら、興味推定の問題においても、紐付けの問題を解決することで、実際にアンケートなどで聞いた回答でログを補完することが可能になり、高精度推定結果を期待できる。

本稿では、クラウドソーシングを用いたアンケート結果と、検索ログを組み合わせた計算社会科学調査のプラットフォームを提案する。 クラウドソーシングとは、ユーザに少額の報酬でアンケートを実施するサービスである。 具体的には、Yahoo! JAPAN が持つマイクロタスク型のクラウドソーシングサービス (http://crowdsourcing.yahoo.co.jp/) を利用して興味推定のアンケート調査を行い、ユーザ ID を検索ログと結合することで、[Blumenstock et al., 2015]と同じく、アンケート結果を持たない検索ログのみ入手可能なユーザについても興味を推定する。応用としては広告の推薦を目的とし、特にユーザの興味推定についての検証を行う。　ただし、本プラットフォームの適用範囲は広告に留まらず、計算社会科学一般であり、[Blumenstock et al., 2015]同様、デモグラフィックデータの推定などにも適応可能である。

2 既存研究

計算広告の分野では、オンライン広告の効果を向上するため、Contextual advertising [Ribeiro-Neto et al., 2005]や、Behavioral targeting [Yan et al., 2009]などの手法が使われてきた。　いずれも、デジタルトレースを使うのみで、キーワードによる広告のマッチングでは足りない部分を補うためタキソノミー[Broder et al.,2007]、Wikipedia から抽出したオントロジーの活用[Wu et al]や、LDA を用いたユーザの行動履歴によるクラスタリング[Tang et al., 2011]といったことは行われるが、ユーザに対する直接のアンケートなどは行われない。

しかし、Tumblr での広告システムの例 [Grbovic et al., 2015] を見るように、ユーザの興味に応じて広告を出し分けたいという要求は強く存在する。　Tumblr ではユーザの投稿データを教師データとして利用してユーザの興味による分類を行うが、この段階でアンケート調査が可能であれば、サービスに対して投稿は行わないがデジタルトレースは残す層に対しても機械学習に必要な訓練データの作成が可能になる。　クラウドソーシングを使うことで興味推定の機械学習に利用する訓練データを広告プロダクト毎に動的に作成できるのが本プラットフォームの強みである。

アンケート調査とデジタルトレースを組み合わせ、本稿のように商品広告の推薦システムに応用した既存研究としては、[Liu et al., 2016]が挙げられる。　[Liu et al., 2016]では、ツイッターのユーザに対し、Big five と呼ばれる心理学で有名なパーソナリティ分類のアンケート調査を行う。　こうして得られたパーソナリティの傾向と、ツイッターでの呟きを紐付け、機械学習を行うことで、アンケート調査を行わないユーザに関してもパーソナリティ傾向の推測を可能としている。　こうして得られたパーソナリティ傾向を利用することで、 [Liu et al., 2016]では若干の購買傾向の上昇が観察され、メーリングリストによる商品の紹介にもパーソナリティ毎に文言を変更すると言った実用化が行われている。

[Liu et al., 2016]によるパーソナリティ傾向の利用は、どのようなサービスにも応用可能という意味で汎用性があり、年齢や性別といったサービスへの登録情報と比べればリッチではあるが、特定のサービス、目的に特化したアンケートではなく、大きな購買傾向の上昇はみられな

いことが分かっている。　やはり、本格的な社会科学の調査や広告効果向上には、性別や年令、パーソナリティ傾向などの汎用的なアンケート項目だけではなく、前節で紹介した [Blumenstock et al., 2015] のように目的毎に作り込まれたアンケート調査の実施が必要である。そこで、本稿では、クラウドソーシングによりアンケート結果を逐次動的に入手し、１）アンケート調査による代表性の高いサンプルの選出、２）行動履歴を用いた予測モデルの構築、３）予測モデル改善のフィードバック、４）予測モデルによる欠損データの補完と実用化という４ 段階を繰り返すことによって、業務の継続的な改善を可能にするプラットフォームを提唱し、これを応用して、検索行動の類似性を用いて推薦を行う推薦システムを検証する。

3 提案手法

開発したシステムは、１）アンケート調査による代表性の高いサンプルの選出、２）行動履歴を用いた予測モデルの構築、３）予測モデル改善のフィードバック、４）予測モデルによる欠損データの補完と実用化という４つのステップからなる。以下にそれぞれについて説明する。

また、適応可能なシナリオとして、Yahoo JAPAN が提供するインタレストカテゴリターゲティングという広告商材の最適化配信に当てはめた例を解説する。　インタレストカテゴリターゲティングとは、ユーザのページ閲覧、検索、広告へのクリックといった行動から興味カテゴリを推定し、ユーザが該当ページやサービスに訪問した際に、ユーザの興味に合ったカテゴリの広告を表示するというものである。　今回は実証実験もこのシナリオで行うこととする。今回の実験ではサンプルは全て弊社のサービスのユーザであるため、本稿ではサンプルとユーザが指す対象は同じである。

3.1 アンケート調査による代表性の高いサンプルの選出

今回の調査対象は、調査者の指定する属性をもつユーザとする。　例えば、ダイエットに興味があるユーザが今回の調査対象であるとする。　この調査対象を代表するサンプルの選出のため、まず、調査者がユーザを特定するアンケート設問を考え、それをクラウドソーシングサービスに登録する。　その設問上で、過去の検索クエリを個人が特定されないレベルで解析に使用することの同意も併せて得た上で、クラウドソーシングの回答結果を解析し、調査対象となるユーザをリストアップする。　このようにクラウドソーシングを通じて、限られたクラウドソーシングのワーカから、調査者が指定する属性を持つユーザ群を抽出するのが最初のステップである。　調査者は事前に質問を練り、クラウドソーシングのユーザ群から代表性の高いサンプルを選出する方法を十分に考えておく必要があり、この段階で通常のアンケート調査同様の試行錯誤が発生する。

今回の実証実験は Yahoo Japan が提供するインタレストカテゴリマッチング広告における６つのカテゴリにおいて行った。６つのカテゴリは、「ダイエット、髪、スキンケア、美容コスメ、レディスファッション、ゲーム」である。　比較的提供する広告数も多いカテゴリを採用した。 代表性の高いサンプルの選定プロセスでは、同じく Yahoo! JAPAN が提供するヤフークラウドソーシングサービスのユーザに質問を割り当てる。　各カテゴリに対し、興味の程度やそれに伴う活動の有無（直近数週間での検索などの情報収集や購買、クーポンの利用の有無な

ど）といった質問を尋ね、これらの設問に対する回答を一週間で約３５００人のクラウドソーシングユーザから得た。　こうして得た回答から、カテゴリに強い興味を持つユーザを絞り込み、その結果、代表性の高いサンプルを選定した。　注意点としては、後段の予測モデルの構築に役立つサンプルが必要であるため、選定したユーザ数が１００人を下回るなど、少なくなりすぎないこと、行動履歴を残しているユーザを選定することが挙げられる。

3.2 行動履歴を用いた予測モデルの構築

次にクラウドソーシングにより選出された代表性の高いサンプルと行動履歴を自動的に紐付けることで、行動履歴から属性を予測する機械学習のモデルを構築する。　機械学習のモデルは様々なものが考えられるが、今回は学習結果の解釈が可能な標準的な分類器としてロジスティック回帰を利用したものとして解説し、実証実験もロジスティック回帰で行う。　分類器を学習するにあたり、代表性の高いサンプルを正例、代表性の高いサンプル以外の一般ユーザ（クラウドソーシングユーザ）を負例とし、行動履歴として検索クエリを素性に利用することで予測モデルを構築する。　これにより、対象カテゴリに興味を深く有するユーザを判定するモデルファイルを得、各特徴量が代表性の高いサンプルとそうでない一般のユーザとを分ける度合いをスコア付きで算出する。 線形モデルであるため、こうして学習した素性の重みから代表性の高いサンプルに特徴的な検索クエリの特徴を抽出が可能となる。

注意点としては、負例にクラウドソーシングを利用しない一般ユーザを利用してしまうと、正例にはクラウドソーシングサービスのユーザだけが含まれてしまい、クラウドソーシングのユーザに特有の特徴を学習器が選んでしまうことである。　負例として、代表性の高いサンプル以外の一般ユーザをクラウドソーシングユーザのログから選定することで、こういったユーザ特徴を相殺し、興味カテゴリに関連する特徴的な検索クエリが発見できる。　なお、実証実験では、より両者の特徴を際立たせる目的で、正例と負例の数のバランスを揃えるようにした。　圧倒的に正例の数が負例に比べてすくなくなるため、正例は全員採用し、負例については正例と数が等しくなるようにランダムで抽出した。　機械学習に用いるハイパーパラメータはクロスバリデーションの評価を行うことで最適なパラメータを選択した。

3.3 予測モデル改善のフィードバック

本提案手法の特徴として、予測モデル改善のフィードバックがある。　フィードバックは代表性の高いサンプルを選定するプロセスおよび行動履歴を用いた予測モデルの構築の２つの段階でかけることが可能である。

まず代表性の高いサンプルの選定プロセスにおけるフィードバック機能について説明する。たとえば、絞り込まれた代表性の高いサンプルの数が理論値とくらべて多すぎる場合は、よりユーザを絞り込む設問を追加で行う。　逆に少なすぎる場合は、設問の条件を緩和する。　あるいは、集まった自由回答のコメントを読み、選定されたユーザが本当に代表性の高いサンプルのユーザとしてふさわしいかを確認し、ふさわしくない場合には設問を調整して、再度ユーザを集めるというステップが考えられる。

次に行動履歴を用いた予測モデルの構築の段階におけるフィードバック機能について説明す

る。このモジュールではモデルファイルがアウトプットとして導出できる。　そのモデルファイルを人が確認し、実際にターゲットとしたいユーザの特徴をある程度表しているかどうか、を確認する。　確認し、想定からずれている場合は再度最初のステップに戻り追加の設問やアンケートの取り直しなどを行う。

これらのフィードバックをうまく活用することで、実際に実用化する以前の段階で人の知識を使った予測モデルの作成が可能となる。

3.4 予測モデルによる欠損データの補完と実用化

こうして得られた予測モデルに一般ユーザの行動履歴を当てはめることで、一般ユーザに欠損している属性データの補完が可能となる。　例えば、アンケート調査でニート（若年無業者）予備軍の代表性の高いサンプルを得られれば、ニート予備軍か否かという属性の欠損データを補完することが可能になる。　弊社の一般ユーザで同様の兆候を示すユーザの数や割合を知ることができれば、計算社会科学分野や、行政における政策決定において有用であると考えられる。

実証実験では、最後に作成されたモデルファイルを読み込み、広告配信ページに訪れた一般ユーザの興味（上記の「ダイエット、髪、スキンケア、美容コスメ、レディスファッション、ゲーム」の６つのカテゴリ）に最適な広告を配信するライブテストを行った。　ユーザが広告配信ページに訪れた際、当該ユーザの過去の検索クエリをデータベースから参照し、ユーザの検索クエリの中でモデルファイルに含まれている特徴量の重みの総和を、ユーザのスコアとして計算する。　このスコアはユーザの対象カテゴリに対する興味の度合いとみなすことも可能である。　ユーザのスコアが閾値より超えた場合に、優先的に対象カテゴリの広告を配信し、もし、興味が閾値を超えていても当該カテゴリに表示する広告が無い場合は、通常の広告配信の興味推定ロジックにより選定された広告が配信されることとした。

4 実証実験

実験は、提案手法の項で述べた Yahoo JAPAN が提供するインタレストカテゴリターゲティングという広告商材の最適化配信のシナリオで行った。　以下に提案手法の項で述べられなかった比較手法、実験設定についての詳細を述べた後、実験結果について解説する。

4.1 比較手法

比較手法は、2 つの方法を準備した。

・rule-based アプローチ（baseline）　ルールベースアプローチとして、人が対応付けた単語と広告カテゴリの組み合わせを準備する。 さらにユーザの検索行動や閲覧ページから広告カテゴリの単語の量をカウントし、一定以上の閾値を超えた場合に、対象ユーザの興味として当該カテゴリを付与する方法である。 たとえば、事前に「ハワイ」という単語と、「海外旅行」という広告カテゴリを人の知識を使って紐付ける。ユーザが「ハワイ」と 5 回以上検索したらユーザに海外旅行の興味を 付与するという仕組みである。閾値は過去のデータから判断し、シミュレ

ーション等を使い機械的に決定されるが、単語と広告カテゴリのヒモ付は人の知識で行う。 この手法を baseline とする。

・click-based アプローチ(click)　代表性の高いサンプルを選定するプロセスでクラウドソーシングを用いずに、クリックの回数 で判断して行う。各カテゴリの広告を過去にクリックした回数をカウントし、それの多いユーザを代表性の高いサンプルとして想定する。 その後のプロセスには同じプロセスで行う。　クリックの回数ベースで代表性の高いサンプルを選定しているため、人手による処理は介することなく、機械的にユーザの興味を付与できるというメリットを持つ。　一方でクラウドソーシングを用いる提案手法のように詳細なターゲット選定はできなくなる。

4.2 実験設定

ライブテストの期間は 2 週間とした。　ユーザそれぞれにつき、比較手法の項で述べた閾値を超えた場合に興味があるとして、ユーザがページを訪れた際の広告カテゴリの引当てを行い、そのカテゴリの広告を取得した。　ユーザに興味ワードを付与する方法以外（時期や提供する広告の内容）は、完全に同一の条件としている。　したがって、結果に差異が発生するとすれば、それはユーザのカテゴリに対する興味の度合いを計測する方法のみといえる。　なお、click-base や提案手法にはモデルを学習する期間が必要となるため、広告提供の 3 週間前から 2 週間のログを持って学習した。

4.3 実験結果

実験の評価は Click Through Rate (CTR)によって行う。CTR は以下の式で定義されるもので、表示された広告うちの何割がクリックされたのかを意味する指標である。

CTR＝クリック数÷表示回数（%）

なお、守秘義務の都合上、CTR の数値自体は掲載できないため、baseline を 0 %としたときの手法間の比較の上昇率/下降率のみを結果として示す。　ライブテストの結果、click ベースのモデルが baseline と比較して 6%程度の伸び率を示した。　その一方、提案手法では baseline に対して 28%程度の伸び率を実現していることが判明した。　この結果、クラウドソーシングを利用して質の高い属性データから予測モデルを構築するほうが、予測精度に優れていることが分かる。

5 結論

本稿ではクラウドソーシングを用いたアンケート結果と、検索ログを組み合わせた計算社会科学調査のプラットフォームを提案し、ユーザの興味の推定が必要な広告配信のシナリオで実証実験を行った。　結果として、click-base の手法と比べ、アンケート結果を利用する方が高い精度で広告配信を推薦することが可能であることが判明し、本プラットフォームの有用性を示す結果が得られた。　さらに、本プラットフォームの適用範囲は興味推定にかぎらず、デモグラフィックデータの推定などにも応用可能であり、社会科学分野での課題解決へ向けた活用が有望視されている。　例えば、アンケート調査でニート（若年無業者）予備軍の代表性の高いサ

ンプルを得て、弊社の一般ユーザで同様の兆候を示すユーザの数や割合を知ることができれば、支援政策の決定において非常に有用であると考えられる。　計算社会科学分野において、本提案プラットフォームの活用が進み、社会の課題解決へとつながることが期待される。

謝　辞

本研究を進めるにあたり、ご協力を頂いた Yahoo! クラウドソーシング、データ＆サイエンスソリューション本部と研究所の皆様に感謝致します。

参考文献

[Blumenstock *et al.*, 2015] Joshua Cadamuro, and Robert On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073– 1076, 2015.

[Broder *et al.*,2007] Andrei Broder, Marcus Fontoura,Vanja Josifovski, and Lance Riedel. A semantic approach to contextual advertising. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 559–566, New York, NY, USA, 2007. ACM.

[Cheng *et al.*, 2012] Haibin Cheng, Roelof van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou, and Vidhya Navalpakkam. Multimedia features for click prediction of new ads in display advertising. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 777–785, New York, NY, USA, 2012. ACM.

[Grbovic *et al.*, 2015] Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, and Ananth Nagarajan. Gender and interest targeting for sponsored post advertising at tumblr. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1819–1828, New York, NY, USA, 2015. ACM.

[Liu *et al.*, 2016] Zhe Liu, Yi Wang, Jalal Mahmud, Rama Akkiraju, Jerald Schoudt, Anbang Xu, and Bryan Donovan. To buy or not to buy? understanding the role of personality traits in predicting consumer behaviors. In *Social Informatics - 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part II*, pages 337–346, 2016.

[Ribeiro-Neto *et al.*, 2005] Berthier Ribeiro-Neto, Marco Cristo, Paulo B. Golgher, and Edleno Silva de Moura. Impedance coupling in content-targeted advertising. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 496–503, New York, NY, USA, 2005. ACM.

[Tang *et al.*, 2011] Jian Tang, Ning Liu, Jun Yan, Yelong Shen, Shaodan Guo, Bin Gao, Shuicheng Yan, and Ming Zhang. Learning to rank audience for behavioral targeting in display ads. ACM, January 2011.

[Wu *et al.*, 2011] Zongda Wu, Guandong Xu, Rong Pan, Yanchun Zhang, Zhiwen Hu, and Jianfeng Lu. Leveraging wikipedia concept and category information to enhance contextual advertising. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 2105–2108, New York, NY, USA, 2011. ACM.

[Yan *et al.*, 2009] Jun Yan, Ning Liu, Gang Wang, Wen Zhang, Yun Jiang, and Zheng Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 261–270, New York, NY, USA, 2009. ACM.

略歴

著者 1 氏名：清水伸幸　2007 年より東京大学情報基盤センター特任助教。2010 年より同センター特任講師。2011 年より Yahoo! JAPAN 研究所上席研究員。クラウドソーシングと計算社会科学、自然言語/画像処理と人工知能の研究開発に従事。博士 (情報工学)。

著者 2 氏名：坪内孝太　2012 年 3 月まで東京大学でオンデマンド交通システムの研究に従事。2012 年 4 月より Yahoo! JAPAN 研究所上席研究員、データサイエンティスト。Yahoo! JAPAN 研究所では人の行動ログ（位置情報、検索ログ、買い物履歴、センサーデータなど）に着目したデータ解析の研究に従事している。2010 年に東京大学にて博士（環境学）の学位を取得。