

Q&A サイトでの閲覧者の回答確率予測とオンライン質問推薦

東京工業大学 *星野雄毅 HOSHINO Yuki
 東京工業大学 田崎誠 TASAKI Makoto
 東京工業大学 水谷圭佑 MIZUTANI Keisuke
 東京工業大学 浅見元哉 AZAMI Motoya
 東京工業大学 石塚湖太 ISHIZUKA Kota
 01405430 東京工業大学 中田和秀 NAKATA Kazuhide

1. はじめに

Web 上の Q&A サイトでは情報を能動的に手に入れることができ、その便利さから多くの人に利用されている。一方で、次のような課題点も存在する。まず、質問者側の視点での課題点として、質問に対して回答が見つからない場合が多いことが挙げられる。このことは、質問者が Q&A サイトに対し興味を失い、サービスからの離脱率を高めてしまう原因となっている。次に、閲覧者側の視点での課題点として、数多くの質問が投稿されている Q&A サイトの中で、閲覧者が答えたい質問を見つけるのが難しいことが挙げられる。そのため、閲覧者が回答をする機会を失っている。

以上の課題点を解決するために、本研究では機械学習と最適化を組み合わせたモデルを提案する。そこでは、まず回答情報を考慮した質問文情報の効率的な抽出方法について考案し、それを用いてサイト閲覧者の各質問に対する回答確率を予測する。次に、予測した回答確率と質問者の満足度を考慮した上で、閲覧者に回答しそうな質問を推薦する。

2. 関連研究

自然言語の埋め込み手法とオンライン最適化の二つの観点から関連研究を紹介し、本研究との関係について述べる。

2.1. 自然言語の埋め込み

自然言語情報の埋め込みを行う手法として BERT [2] が一般的に使用されている。現在公開されている事前学習済み BERT モデルは Wikipedia やニュースデータを用いて学習している。このため、これらのモデルを本研究が対象とする育児 Q&A サービスに転用したときの精度は、絵文字や育児

に固有の言葉に対応できないことから十分でない可能性がある。

2.2. オンライン最適化

(オフライン) 最適化が全データに基づき最適な解を算出するのにに対し、オンライン最適化は逐次的に入力されるデータに対して最適な解を算出するものである。特に、Q&A サイトでのオンライン最適化として、質問が投稿されるたび、質問に対し過去の回答実績に応じて上位にランクされた回答者を逐次的に推薦する多目的最適化問題を考えた研究が存在する [1]。一方で、多くの質問に回答がばらつくように推薦を最適化する研究は進んでいない。

3. 提案手法

提案手法は回答確率を予測する回答確率予測モデルと、その予測に基づいたオンライン最適化という 2 つの部分からなる。

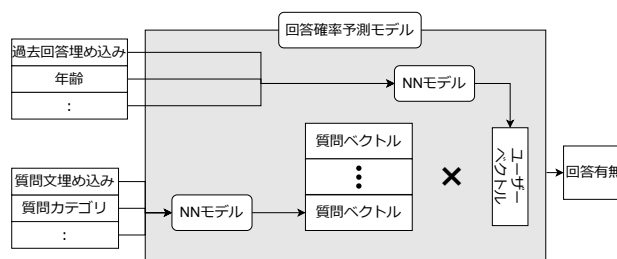


図 1: 回答確率予測モデルの全体像

回答確率予測モデルはニューラルネットワークを用いて構築しており、その全体像を図 1 で示した。これは、閲覧者のデモグラフィック属性と投稿された質問文を特徴量として、特定の閲覧者が特定の質問に対して回答する確率を予測する。ただし、質問文の自然言語情報をそのまま入力するのは大規模なモデルが必要となるため、埋め込みベ

クトルの状態で入力を行った。この際、事前学習済みのBERTを利用して埋め込みベクトルを獲得するのではなく、絵文字や育児固有の表現に対応するためにBERTの事前学習を行った。事前学習は質問回答データおよそ500万件のデータを使用し、TeslaO199(5.3Tflops)を128台用いて並列処理して5日間かった。さらに、予測精度を向上させるため、回答の傾向を考慮できるようにsimcse[3]をベースとしたファインチューニング手法を提案している。

次に、閲覧者に対する質問の推薦手法について説明する。本研究では、閲覧者が対象のサイトを閲覧した際にすぐ推薦結果を表示することを目指すため、全体最適化でなく逐次的な意思決定であるオンライン最適化を行う。この最適化において、質問への回答数に基づいた質問者の満足度を考慮している。一般的な推薦システムでは回答確率が最大の質問を推薦するが、これでは回答しやすい質問に偏った推薦が行われることが多い。すると、回答が付きにくい質問に対して推薦が行われなくなり、無回答となった質問の質問者のユーザー体験を損ねてしまうと考えられるからである。そこで、回答数を与えると質問者の満足度を導く単調増加な凹関数を考え、全質問者の満足度の和を最大化するように推薦を行う。本研究では、この関数として回答数 N に対し $\log(1+N)$ を用いている。また、オンライン最適化の時点では将来得られる回答数は分からないため、LightGBM[4]によって予測した値を用いて補完し、満足度増加量の期待値が最大となる質問を推薦する。

4. 実験

提案手法について、予測精度の改善と最適化の効果を検証するための数値実験を行った。ここでは、後の最適化による効果を測ったシミュレーションの結果を載せる。データは株式会社コネヒト様から頂いたママリというQ&Aサービスのデータを利用した。2021年1月から6月までのおよそ100万質問を学習データとし、7月1日から7日までの4万件の質問をテストデータとした。

表1がシミュレーションによって得られた満足度の上昇率である。これを見ると、回答確率が最大のものを推薦する場合と比較して、最適化を加えたものは安定的に満足度が高いことが分かる。

表 1: 満足度関数を用いたシミュレーション結果

回答数 / 月	1000 万	3000 万	5000 万
回答確率最大	5.21%	7.31%	4.82%
過去回答のみ	8.79%	17.18%	21.65%
提案手法	10.09%	20.57%	25.17%

また、過去回答数のみを考慮した推薦と比較して、提案手法は将来の回答数予測も含めて最適な推薦を行うことで、満足度が向上していることも分かる。なお、追加実験によって、単純に回答確率が最も高いものを推薦してしまうと、推薦数が増えたときに一部の質問に回答が集中してしまい、全体のユーザー満足度が下がってしまうことが確認できた。

謝辞

本研究は、経営科学系研究部会連合協議会-データ解析コンペティション-の一環として行われました。主催者の皆様に感謝申し上げます。また、本研究をするにあたりまして、データを提供して下さいコネヒト株式会社様にも感謝申し上げます。

参考文献

- [1] Xiang Cheng, Shuguang Zhu, Sen Su, and Gang Chen. A multi-objective optimization approach for question routing in community question answering services. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, No. 9, pp. 1779–1792, 2017.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [3] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.
- [4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, Vol. 30, , 2017.