専用アプリ上の質問データに基づく子育てライフステージの課題変化分析に関する研究

1X18C093-9　山田晃輝
指導教員　後藤正幸

1. 研究背景・目的

乳児の子育てを行う親には，様々な悩みが存在する．こうした悩みを解消するため，オンライン上に子育て特化型の質問投稿サービス（以下 A サービス）が展開されている．A サービスでは，ユーザが投稿した子育てに関する質問に他のユーザが回答することで，質問者や閲覧者は子育てに関する知見を得ることができる．ここで，質問内容にはユーザの悩みが反映されており，ユーザの悩みは子どもの成長度合い，すなわちライフステージに応じて変化すると考えられる．そのため，質問データからユーザの悩みの推移を子どものライフステージごとに捉えることができれば，ユーザ毎に適切なタイミングで適切な情報を提供するなど A サービスの利便性向上に大きく寄与することが期待できる．

以上のような質問データから，親の悩みの内容を分析する手法としてトピックモデルがある．これをテレビ視聴履歴の時間的なトピック推移分析に援用し，可視化する手法として，坂元ら [1] の手法が挙げられる．坂元らの手法では，視聴がない期間は未視聴として扱うことができた．しかし，A サービスの質問データの場合，質問投稿がない時期（非質問期間）でも子育ての悩みは生じており，一律に未質問期間を悩み無しとしてまとめてしまうのは不適切である．そのため，本研究では非質問期間における悩みの推定・予測が必要になる．

そこで本研究では，質問データを 1 質問 1 トピックで捉えることで，単一のトピックから成ると想定される質問の内容を表現し，そこから生後 48 週以内の乳児のライフステージの変化に伴うユーザのトピック推移を捉える手法を提案する．さらに，得たユーザのトピック推移を確率的に表現することで，ユーザのある時期のトピックから，その次の非質問期間に遷移する可能性が最も高いトピックを推定する手法を構築する．これにより，ユーザごとに次に遷移すると推定されたトピックに関する記事を前もって提示するなど，先回りした支援・施策が可能になる．実データ分析では，提案手法を適用した結果を示し，トピック推定精度の観点から提案手法の有用性を示すために評価実験を行う．

2. 従来手法

ユーザの時系列に沿ったトピック推移を捉える手法として，坂元らの研究が挙げられる．坂元らは，テレビドラマの視聴データから視聴者の視聴傾向を時期ごとにクラスタリングし，その結果を量的な流れを表現する図であるサンキーダイアグラムを用いて可視化した．具体的には，初めに全体の視聴履歴から番組のジャンルを表すクラスタを学習し，次にクラスタの意味合いを保持しながら期間ごとに各視聴者をクラスタに割り当てている．

3. 提案手法

坂元らの手法では，1 人の視聴者に複数トピックを想定していた．しかし，本研究で対象とする質問データは，短文であるため 1 質問当たり 1 トピックを想定するのが自然である．そのため本研究では，質問データを 1 質問 1 トピックで捉えることで，単一のトピックから成ると想定される質問の内容を表現し，トピック推移を捉えることを考える．それに加え，得られたトピック推移を確率的に捉えることで，ユーザのある時期のトピックから，その次の非質問期間に遷移する可能性が最も高いトピックを推定する手法を提案する．提案手法を，以下の全 4 ステップで示す．

Step.1　混合ユニグラムモデル [2](Mixture of Unigram Models，以下 MUM）を用い，全期間の質問データからトピック内単語分布 $\boldsymbol{\Phi}$ を学習する．

Step.2　ユーザごとに期間を区切り，再度 MUM を用いてトピック分布 $\boldsymbol{\theta}$ を学習する．このときトピック内単語分布 $\boldsymbol{\Phi}$ は Step.1 の推定値を用いる．

Step.3　[可視化分析] サンキーダイアグラムを用いて，ユーザごとのトピック推移をまとめて可視化する．

Step.4　[トピック推定・予測] 得られたトピック推移を確率的に捉え，ユーザに対し次の非質問期間に遷移する可能性が最も高いトピックを推定する．

3.1. トピック内単語分布の学習

Step.1 では，MUM を用いて全期間の質問データを学習する．MUM のグラフィカルモデルは，図 1 で表される．

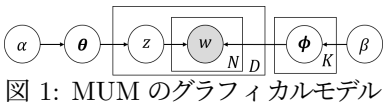

図 1: MUM のグラフィカルモデル

ここで，α，β はハイパーパラメータ，$\boldsymbol{\theta}$ はトピック分布，$\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_K)$ はトピック内単語分布，z は各文書のトピックを表す．また，N_d は質問 d における単語数，D は質問数，K はトピック数を表す．パラメータの推定は，崩壊型ギブスサンプリングを用いて行う．

3.2. 期間ごとのトピック割り当て

ここでは，ユーザ $u_i (i = 1, ..., I)$ ごとに，期間 $l (1, ..., L)$ を区切って再度 MUM の学習を行う．Step.2 におけるユーザごとの MUM 適用のイメージを図 2 に示す．

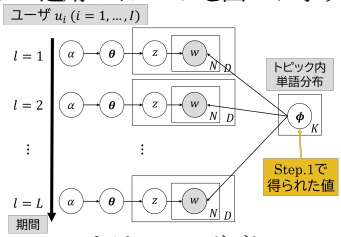

図 2: Step.2 におけるユーザごとの MUM 適用

パラメータの学習は，Step.1 と同じく崩壊型ギブスサンプリングで行う．ここで，$\boldsymbol{\Phi}$ は Step.1 で得られた値で固定し，更新を行わない．ユーザ u_i の期間 l において割り当てるトピック c_{il} は，以下の式 (1) で表される．

$$c_{il} = \begin{cases} K+1 & \text{if } D_{il} = 0 \\ \arg\max_k \theta_{ilk} & \text{otherwise} \end{cases} \quad (1)$$

ここで，D_{il} はユーザ u_i が期間 l に投稿した質問数を表す．また，$K+1$ は非質問期間に割り当てるトピックである．

3.3. トピック推移の可視化

Step.3 では得られたユーザごとのトピック推移をサンキーダイアグラムを用いて可視化する．なお，非質問期間の推移は含めず，投稿質問がある期間の推移のみ図に含める．

3.4. 次のトピック推移の推定

Step.4 では，得られたトピック推移を確率的に捉えることで，ユーザに対し，次の非質問期間に最も遷移する可能性が高いトピックを推定する．ここで，期間 l においてトピック j であったユーザに対し，トピック j から最も多くの人数が遷移している先のトピック，すなわち $\arg\max_k N^l_{jk}$ を期間 $l+1$ の推移先トピックと推定する．ここで，N^l_{jk} は，期間 l にトピック j に所属していたユーザのうち，期間 $l+1$ にトピック k に推移した人数を表す．

4. 実データ分析

4.1. 分析条件

本研究では，経営科学系研究部会連合協議会主催，令和 3 年度データ解析コンペティションで提供された A サービスにおける質問データおよび子ども情報のデータを用いる．データの対象期間は 2019 年 1 月 1 日から 2021 年 7 月 31 日までである．質問データには，質問を投稿したユーザの ID，カテゴリ ID，質問本文，質問投稿時間が含まれる．また，子ども情報のデータには，ユーザ ID と登録された子どもの出産日（出産前の場合は出産予定日）が含まれる．

実データ分析で対象とする質問データは，投稿日が第 1 子の生後 48 週以内である 1,092,111 投稿とする．また，トピック数 K を 10，期間数 L を 12(1 期間は 4 週）とする．Step.1 では，全データを用いてトピック内単語分布の学習を行った．また，Step.2 以降では質問数が 100 以上のユーザ 1,928 人について，トピック推移を推定した．

4.2. 分析結果

表 1 に，Step.1 で得られたトピックのうち，可視化時に重要となる主要なトピックとその主な出現単語を示す．

表 1: 得られた主要なトピックと主な出現単語

トピック番号	主な出現単語
1（妊娠・出産）	生理，産後，陣痛，妊娠
6（ミルク）	ミルク，母乳，授乳，おっぱい
7（人間関係）	旦那，実家，義母，夫
8（食事）	離乳食，食，野菜，粥
9（子どもの世話）	抱っこ，昼寝，寝返り，泣き

次に，Step.2 で得られたトピック推移を可視化したサンキーダイアグラムを図 3 に示す．図 3 より，初めはトピック 1（妊娠・出産）に多くのユーザがいるが，8 週目以降はトピック 6（ミルク）に所属するユーザが多くなることが分かる．また，24 週からはトピック 8（食事）に興味のあるユーザが増え始め，28 週からは多くのユーザがトピック 8（食事）に興味を持続的に持つようになる．

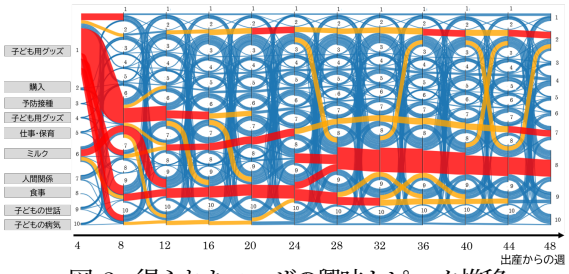

図 3: 得られたユーザの興味トピック推移

4.3. 評価実験

ここでは，提案手法の推定精度を評価するため，全期間に少なくとも 1 つ以上投稿が存在するユーザ 924 人について，実データを用いて次のトピック推移を推定し，その推定精度を他の手法と比較する．比較手法として，推移先のトピックをランダムに選択するランダム法と，その期間に所属する人数が最も多いトピックを選択する最多所属法を用いた．実験結果を図 4 に示す．

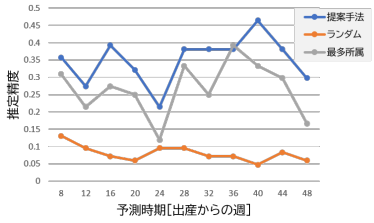

図 4: 推定精度の比較

図 4 より，提案手法はいずれの比較手法よりも推定精度が高いことが示された．一方で，最多所属法と比較すると，出産後 36 週のみわずかに提案手法の推定精度が劣っている．ユーザのトピック所属状態を図 3 より確認すると，この期間はトピック 8（食事）に多くの割合のユーザが所属していることが分かる．よって，提案手法はユーザの遷移が多様である場合に良い推定精度を示すことが示唆される．

5. 考察

実データ分析より，提案手法を用いることで，ユーザの悩みの推移を可視化し，かつ次のトピック推移を高い精度で推定することができた．このことから，例えば提案手法で次のトピック推移が「離乳食」と推定されたユーザに対し，離乳食についての解説記事を提示するなど，先回りした的確な支援に提案手法が大きく貢献できると考えられる．

6. まとめと今後の課題

本研究では，A サービスにおける質問データから，ユーザの興味トピックの推移を推定するとともに，確率的に今後ユーザが遷移するトピックを推定する手法を提案した．さらに実データに適用して，質問データから得られる分析結果と様々な知見を示した．今後の課題としては，異なる期間や別のデータへの適用が挙げられる．

謝辞：本研究では，コネヒト株式会社より「乳幼児を持つ親向けポータルサイトの Q&A データ」をご提供頂きました．貴重なデータの提供に深く感謝いたします．

参考文献

[1] 坂元哲平，小林佑輔，中川慶一郎，生田目崇，後藤正幸，“トピックモデルを用いたテレビ視聴におけるトレンド分析方法の提案，” 情報処理学会論文誌，vol. 61, no. 1 , pp. 346–356, 2021.

[2] 岩田具治，“トピックモデル，” 講談社，2015.