Vol.2020-CSEC-91 No.12
Vol.2020-SPT-40 No.12
Vol.2020-EIP-90 No.12
2020/11/25

Q&A コミュニティの質問文を用いた育児の悩みの可視化

吉見 憲二[†1] 谷本 和也[†2] 田中 康裕[†3] 岩井 憲一[†4]
上田 祥二[†5] 針尾 大嗣[†6] 小舘 亮之[†7]

概要：Yahoo!知恵袋に代表される Q&A コミュニティでは，投稿者の悩みが質問というかたちで蓄積されている．こうしたテキストデータは多くの情報量を有するものの，構造化されていないことからなかなか分析が進んでいない現状がある．本研究では，特に投稿者の悩みが多岐に亘る育児関係のカテゴリを対象に，テキストマイニングの手法を用いて「育児の悩み」を可視化し，その関係性について考察する．

キーワード：Q&A コミュニティ，育児の悩み，Yahoo!知恵袋，計量テキスト分析

1. はじめに

日本最大の Q&A コミュニティである Yahoo!知恵袋の投稿データは国立情報学研究所の IDR データセット提供サービスによって提供されており，多くの研究者に利用されている[1]．近年では，ソーシャルメディアの投稿データの分析がさまざまな分野で行われているが，大量のデータの取得や投稿のノイズへの対応などの問題があり，精度の高い結果を得るためには困難が伴う．他方で，Q&A コミュニティにおける質問文は，ソーシャルメディアの投稿に比べて，カテゴリ化されている点や投稿者の問題意識がはっきり表れている点で優位性がある．

本研究では，日本最大の Q&A コミュニティである Yahoo!知恵袋の育児カテゴリを対象に，質問文から育児の悩みを可視化することでその活用の可能性を示すことを試みる．

2. 先行研究

2.1 Yahoo!知恵袋の質問文を用いた先行研究

著者らは国立情報学研究所の IDR データセット提供サービスで提供される「Yahoo!知恵袋データ(第 3 版)」を用いて，「大学受験」カテゴリにおける大学名の共起関係の分析，「スマートフォン」カテゴリにおける iPhone と Android の質問文の比較，「国内観光」カテゴリにおける観光地名を含む質問文の特徴の可視化といった研究を行ってきた[2] [3] [4] ．

こうした一連の研究からは，カテゴリ化された質問文を扱うことの利点が確認されており，さらに多くの分野に応用できる可能性が示されている．

2.2 育児に関する先行研究

Yahoo!知恵袋に投稿された育児の悩みを分析した先行研究では，「赤ちゃん」and「寝かしつけ」を検索ワードとしてヒットした母親の質問 253 件を対象とし，「抱っこ」「母乳」「添え乳」といった月齢別の特徴語や質問者の悩みの実態について明らかにしている．加えて，質問文の分析を通して，育児に必要なサポートの在り方についても言及している[5].

当該研究では，「ソーシャルメディアを利用して寝かしつけに関連した悩みを発言し，他の育児経験者の意見をとおして問題解決を図っていこうとする現代の育児中の母親の生の声」「子どもを育てながらひとつひとつ意思決定していかなければならない現代の母親の，専門職には言えない，または相談するチャンスがない発言」として Yahoo!知恵袋の質問文を評価しつつ，1 年分のデータに限定したためデータ数が限られていることを課題として挙げている．

2.3 問題意識

先行研究からは Yahoo!知恵袋の質問文がさまざまな分野に応用できること，国立情報学研究所の IDR データセット提供サービスを利用することで大規模データへのアクセスが容易になることが示されている．他方で，育児に関する先行研究では，手動での質問文取得が行われており，テーマも「寝かしつけ」に限定されている．より大規模なデータを用いれば，育児に関する広範なテーマの分析が可能になることが期待される．

そこで本研究では，IDR データセットを用いて Yahoo!知恵袋における「育児カテゴリ」の質問文について計量テキスト分析の手法を用いて分析し，有用な知見を獲得することを目的とする．

†1 成蹊大学　Seikei University
†2 佛教大学　Bukkyo University
†3 社会データ構造化センター　Center for Social Data Structuring
†4 滋賀大学　Siga University
†5 株式会社セールスフォースドットコム　salesforce.com, Inc
†6 摂南大学　Setsunan University
†7 津田塾大学　Tsuda University

Vol.2020-CSEC-91 No.12
Vol.2020-SPT-40 No.12
Vol.2020-EIP-90 No.12
2020/11/25

3. 分析

3.1 分析に用いるデータ

本研究では，国立情報学研究所が提供する「Yahoo! 知恵袋データ(第 3 版)」の 2019 年度提供版および 2020 年度提供版における質問文を利用した[1]．Yahoo!知恵袋は日本最大の Q&A コミュニティであり，2020 年度提供版における質問数は約 270 万件，回答数は約 838 万件となっている．ただし，収録データは収録期間(2014 年 4 月 1 日－2018 年 3 月 31 日)に投稿され解決した質問の 10%がランダムサンプリングされたものであるため，当該期間の回答全てが含まれるわけではない．しかしながら，手動での取得に比べて比較にならないほど大容量のデータにアクセスできる点は依然として魅力的である．

分析に当たっては，提供されたデータから「>子育てと学校>子育て，出産>子育ての悩み」カテゴリの投稿のみを抽出した．対象となった投稿は 9,113 件となった．

3.2 分析対象

「子育ての悩み」カテゴリでは妊娠中から成人した子どもまで幅広い対象への相談が行われている．本研究では，先行研究との比較の観点から特に乳幼児に対する質問に着眼したため，生後から 2 歳未満の育児に対する質問を分析の対象とした．具体的には，質問文に記載されている月齢や年齢によって明確に生後から 2 歳未満と判断できることを基準として採用した．

しかしながら，質問文には「年齢表記の表記ゆれ（一歳 6 か月，1 歳半，13 ヶ月)」があったこと，「文脈を踏まえた判断（もうすぐ 1 歳)」が必要な投稿が少なからずあったこと，「子ども以外の対象との判別（猫等)」が必要な投稿があったことなどから，機械的な抽出は著しく困難であった．そのため，最終的には分析対象の抽出にあたって目視での判断を余儀なくされた．本稿では「子育ての悩み」カテゴリにおける 9,113 件の投稿のうち，目視での確認によって明確に生後から 2 歳未満と判断できた分の一部である 794 件の投稿を分析対象とした．

3.3 分析方法

分析に当たってはテキストマイニングのフリーソフトウェアである KH Coder（https://khcoder.net/）を使用した．形態素解析は付属の ChaSen（茶筌）を用いているが，後述の頻出上位語の分析から一定の仕様が見られた「完ミ」「断乳」「トイトレ」を強制抽出の対象語とした．

分析の手順として，まず頻出上位単語について概観し，経年的な全体の傾向について対応分析（コレスポンデンス分析）を用いて検討した．次に，特に強い単語の共起関係について共起ネットワーク分析を用いて可視化し，特徴的なテーマについて把握した．最後に，把握したテーマについて，時期ごとの登場頻度の差異をカイ二乗検定で確認した．分析にあたっては，全体の投稿の 1 割以上の投稿に登場している単語を対象とした．さらに，共起ネットワーク分析の描写に当たっては Jaccard 係数 0.2 以上の共起関係を基準とした．

4. 分析結果

4.1 頻出上位単語

頻出上位 30 語の抽出結果を表 1 に示している．名詞では，「赤ちゃん」「生後」「母乳」「離乳食」といった育児に関する単語が見られた．動詞では，「寝る」「泣く」「飲む」といった単語が上位に見られた．こうした結果は育児の悩みを対象とした質問文に見られるものとして妥当ではあるものの，意外性にも欠けている．そこで，次節では育児の時期を踏まえた分析を行い，時期ごとの登場傾向について検討する．

表 1　頻出上位 30 語（1 投稿当たり）

抽出語	登場数	比率	抽出語	登場数	比率
思う	349	44.0%	いい	158	19.9%
赤ちゃん	258	32.5%	母乳	154	19.4%
生後	251	31.6%	息子	152	19.1%
寝る	241	30.4%	出る	150	18.9%
今	224	28.2%	娘	149	18.8%
泣く	218	27.5%	夜	144	18.1%
言う	215	27.1%	歳	143	18.0%
子供	211	26.6%	離乳食	140	17.6%
お願い	207	26.1%	起きる	139	17.5%
子	202	25.4%	少し	139	17.5%
教える	194	24.4%	食べる	139	17.5%
飲む	190	23.9%	前	139	17.5%
時間	185	23.3%	聞く	138	17.4%
ミルク	183	23.0%	授乳	132	16.6%
最近	159	20.0%	質問	124	15.6%

Vol.2020-CSEC-91 No.12
Vol.2020-SPT-40 No.12
Vol.2020-EIP-90 No.12
2020/11/25

4.2 対応分析

今回対象となった 794 件の投稿について，「半年未満」「半年以上 1 歳未満」「1 歳以上 2 歳未満」」の 3 つの時期に分けて対応分析（コレスポンデンス分析）を行った．結果は図 1 の通りである．横軸を示す成分 1 は 77.97%，縦軸を示す成分 2 は 22.03%の寄与率となった．成分 1 は経年的な変化，成分 2 は離乳食の傾向をそれぞれ示していると解釈できる．

経年的な変化として，生後半年未満では，授乳やミルクの話題が中心となり，半年から 1 歳にかけて離乳食の話題が顕著に登場している．さらに，1 歳以降になると，「赤ちゃん」から「息子」「娘」「子」「子供」へと呼称が変わっている．また，「遊ぶ」「食べる」という話題が顕著に増えている．

全体的な傾向として，「ママ」「母」は出てくるが，「父」に関する用語は頻出単語として出てこないこと，「男の子」は出てくるが「女の子」は頻出単語として出てこないことが興味深い点であった．

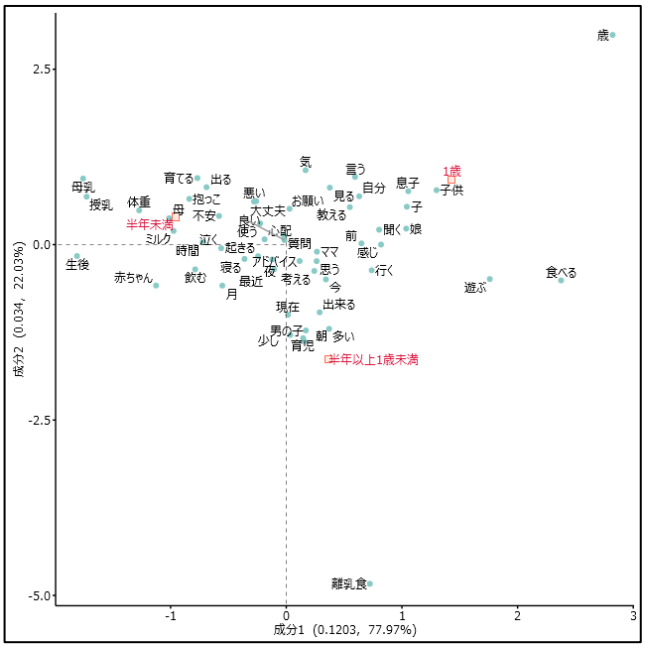

図 1　対応分析（コレスポンデンス分析）の結果

4.3 共起ネットワーク分析

続いて，特に強い単語間の共起関係について，共起ネットワーク分析を用いて可視化した．結果を図 2 に示している．図 2 からは，一般的な用語に加えて，母乳・ミルクに関する話題，寝る/泣くことに関する話題，離乳食に関する話題が見られた．

こうした強い共起関係が見られるテーマは多くの相談者が悩んでいる内容だと考えられる．そこで，時期ごとにこうしたテーマの相談内容がどのように変わるのかについて次節で検討した．

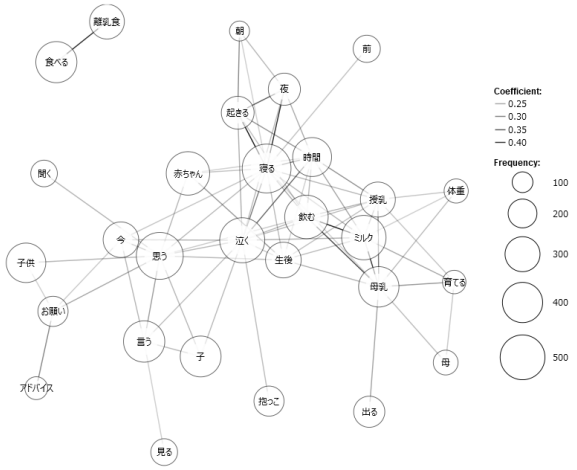

図 2　共起ネットワーク分析の結果

4.4 時期別テーマのカイ二乗検定

共起ネットワーク分析の結果からは，顕著な悩みとして「母乳・ミルク」「寝る/泣く」「離乳食」といったテーマを読み取ることができた．そこで，これらのテーマに関連した単語をグループ化し，時期ごとの登場数の差異についてカイ二乗検定を行った．

各テーマの対象語は表 2 に，カイ二乗検定の結果は表 3 にそれぞれまとめている．表 3 の結果より，各テーマの登場傾向は時期によって異なることが有意に示されている．

表 2　各テーマの対象語

テーマ	対象語
母乳・ミルク	ミルク，母乳，授乳，完ミ，混合，断乳
寝る/泣く	寝る，泣く，夜，昼寝
離乳食	離乳食

表 3　時期別テーマのカイ二乗検定の結果

	母乳・ミルク	夜泣き	離乳食	ケース数
半年未満	183 (51.12%)	197 (55.03%)	27 (7.54%)	358
半年以上 1 歳未満	74 (33.48%)	100 (45.25%)	88 (39.82%)	221
1 歳以上 2 歳未満	44 (20.47%)	79 (36.74%)	25 (11.63%)	215
合計	301 (37.91%)	376 (47.36%)	140 (17.63%)	794
カイ 2 乗値	56.167**	18.557**	105.340**	

*<0.05 **<0.01

Vol.2020-CSEC-91 No.12
Vol.2020-SPT-40 No.12
Vol.2020-EIP-90 No.12
2020/11/25

表 3 より，「母乳・ミルク」に関する話題は，生後半年までの時期に集中し,時間が経つごとに減っていることが読み取れる.「寝る/泣く」に関する話題も同様に，生後半年までの時期に集中し,時間が経つごとに減っている．このことは，生後半年までの大変な時期を乗り切れば，これらの問題はある程度自然に解消されていくものと捉えられるかもしれない．そして，「離乳食」に関する話題も生後半年から 1 年の間に集中しており，1 歳以降には顕著に減少している．

なお，各時期と単語との共起関係を共起ネットワーク分析によって可視化した結果が図 3 であるが，これまでの分析内容と概ね整合的な結果が得られている．

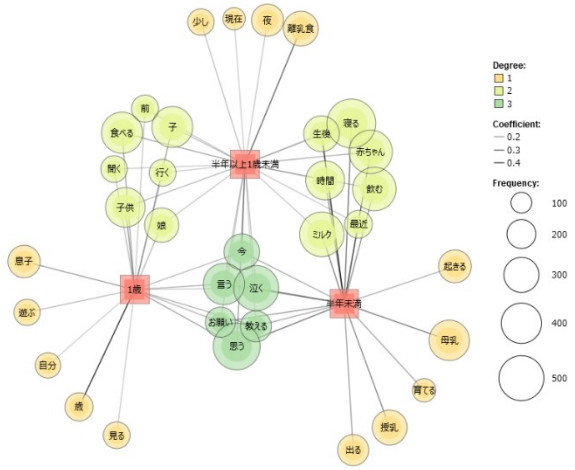

図 3　時期と単語間の共起ネットワーク

4.5 考察

一連の分析を通して, Yahoo!知恵袋の投稿データからある程度蓋然性の高い話題の流れを再現することができた. 結果自体は新規性に乏しいものの，子どもの月齢や年齢によって悩みの質が異なる事実は重要である．

加えて，現在悩んでいる問題がある程度時間の経過によって解決される可能性があることを示すことができたのも本研究の成果だと考えられる．ただし，特に一般的な傾向から逸脱する悩みに関しては，より深刻に捉えるきっかけになってしまうことが懸念されるため，こうした結果の解釈については慎重に扱う必要がある．IDR データセットには質問文だけでなく回答も含まれていることから，回答も含めた分析を行うことでより有用な知見が得られるかもしれない．

また，質問文を扱うにあたって月齢や年齢の表記ゆれが大きな問題となったのも新しい発見であった．IDR データセットを用いたのは大量の投稿を効率的に扱うことが目的であったが，本研究のような分析を行う際には年齢の表記ゆれは深刻な課題となる．この点は，機械的な判別の技術が向上することで解決が期待できるものである．

5. おわりに

本研究では，国立情報学研究所が提供する「Yahoo!知恵袋データ(第 3 版)」の 2019 年度提供版および 2020 年度提供版データを利用し，「子育ての悩み」カテゴリにおける質問文から育児の悩みの分析を行った．「半年未満」「半年以上 1 歳未満」「「1 歳以上 2 歳未満」」の 3 つの時期に分けて行なった分析からは，各時期において表出する悩みが異なること，時間の経過によって自然と解消していく傾向があることが確認された．こうした結果は，特に初めての子育てに悩む親にとって貴重な情報となるかもしれない．

先行研究でも Yahoo!知恵袋の質問文を「子どもを育てながらひとつひとつ意思決定していかなければならない現代の母親の，専門職には言えない，または相談するチャンスがない発言」と表現していたが，こうした生の声を分析することは同じような悩みに直面している多くの親にとって参考になるデータとなることが考えられる．

今後は，回答も含めた分析を行うことで，より現実的な問題に対する有用な知見を提供することを目指したい．

謝辞

本研究は ROIS-DS-JOINT(課題番号:00032,研究代表者:小舘亮之)の助成を受けた.また,国立情報学研究所の IDR データセット提供サービスにより,ヤフー株式会社から提供の「Yahoo!知恵袋データ(第 3 版)」を利用した.

参考文献

[1] 国立情報学研究所：“情報学研究データリポジトリ” (2020/10/26 閲覧), https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html
[2] 吉見憲二（2020）「計量テキスト分析を用いた Q&A コミュニティからの評判情報の推定」『情報処理学会全国大会講演論文集』
[3] 吉見憲二・田中康裕・針尾大嗣・谷本和也・源城かほり・岩井憲一・小舘亮之（2020）「Q&A コミュニティにおける質問文からの製品情報の分析」『情報コミュニケーション学会第 17 回全国大会予稿集』
[4] 吉見憲二（2020）「Q&A コミュニティにおける質問文からの観光情報の分析」『情報処理学会研究報告電子化知的財産・社会基盤（EIP）, 2020-EIP-87(11)』pp. 1-4.
[5] 佐々木裕子，高橋眞理（2015）「インターネットの Q&A コミュニティサイトにみる 0～4 ヵ月児の母親の育児における寝かしつけの悩み　テキストマイニングによる分析」『医療看護研究』11 巻 2 号，pp.28-35.