［論文］

日本最大級の女性向けコミュニティサイトにおける出産予定日予測システムの構築

Prediction System of Estimated Date of Confinement Developed for One of the Largest Woman CQA in Japan

島田 達朗†, 櫻井 彰人‡
Tatsuro SHIMADA, Akito SAKURAI

†Connehito株式会社
‡慶應義塾大学
†Connehito, Inc.
‡Keio University

要旨

多くのオンラインコミュニティサイトにはユーザー登録機能がある．ユーザーが入力をした情報を用いることによって，ユーザー自身が求める情報コンテンツを提供することが可能になり，ユーザーの満足度を向上させる施策により長くサービスを利用してもらうことに繋げることができる．また，パーソナライズされた広告の配信によるサービスの収益最大化が可能になる．特に結婚や出産と言った，人の生涯で発生する重要な出来事（ライフイベント）を起点として，単価の高い購買活動を行うことが知られている．しかし，ユーザーがこのようなライフイベント情報を入力するには一定のハードルがあり，ユーザーによっては自分自身のパーソナルな情報をほとんど入力しない場合も存在する．妊娠・出産・子育ての疑問を解決する女性のためのコミュニティサイトであるママリでも出産日を入力せずにサービスを利用しているユーザーが存在している．そこで，出産日を入力せずとも，ユーザーが投稿した質問から機械学習を用いてユーザーの出産予定日を予測するシステムの構築に取り組んだ．

Abstract

Many online community question answering sites have a user registration function. By using the information entered by the user, it becomes possible for the user to provide the information contents he / she desires, and it is possible to lead the user to use the service for a long time by functions for improving their satisfaction. In addition, it is possible to maximize revenue of services by delivering personalized advertisements. Information on important events (life events) occurring in a person's lifetime, particularly marriage or childbirth, is known as the time when the person collects various information to make decisions and tend to carry out purchase activities of high unit price. However, there are certain barriers for the user to input such life event information, and there are cases in which some users do not input their own personal information. There are users who use the service without inputting estimated date of confinement even at Mamari which is a community question answering site for women who solves the question of pregnancy, childbirth, and child rearing. Therefore, we constructed a system that predicts user's baby's estimated date of confinement by machine learning from the question posted by the user without inputting estimated date of confinement.

1. はじめに

1.1. サービス規模

妊娠・出産・子育ての疑問を解決する女性のための Q&A アプリ ママリ [1] は日本で最大級の女性向け Community Question Answering（以下，CQA と略す）サイトである．CQA サイトとはユーザーの投稿した質問に対して，他のユーザーが回答を寄せるサイトのことを言う．日本においては，Yahoo!知恵袋に代表され

［論文］2017年10月20日受付，2018年1月17日改訂，2018年2月14日受理

る.

2016 年末時点において，ママリへユーザー登録を行い，かつ子供情報を入力し，2016 年に出産を迎えたユーザーの数は 15 万人以上である．厚生労働省発表の人口動態統計 [2] から算出すると 2016 年に出産予定日を迎えた人の約 6 人に 1 人以上がママリに登録していることになり，ママリは国内最大級の規模のサービスとなっている.

1.2. 出産日情報の活用

ユーザーが入力をした情報を用いることにより，ユーザー自身が求める情報コンテンツを提供することが可能になる．その結果，ユーザーの満足度を向上させる施策やパーソナライズされた広告の配信によるサービスの収益最大化が可能になる.

特に結婚や出産と言った，人の生涯で発生する重要な出来事（ライフイベント）を起点として，単価の高い購買活動を行うことが知られている.

広告配信の例として，例えば出産したことがわかれば生命保険や学資保険に関する広告を配信するということが考えられる.

また，出産からある程度時間が経過したユーザーには幼児教育に関するコンテンツを提供することでユーザーの体験を向上させ，より長くサービスを利用してもらうことに繋げることができる．ママリでは妊娠して経過した月数や週数によって，コンテンツのレコメンデーションを行っている．具体的には図 1 のような画面である．このように出産日情報には多くの活用先がある.

図 1 妊娠週数に合わせたコンテンツレコメンデーションの例

1.3. サービス課題

ユーザー数が増加する一方でユーザーが出産予定日のようなライフイベント情報を入力するには一定の精

神的なハードルがある．昨今の個人情報に対する警戒感の高まりもあって，パーソナルな情報をほとんど入力しないユーザーも存在する．ママリでも出産日を入力せずにサービスを利用しているユーザーが約半数以上存在している．

そこで，出産日を入力せずとも，ユーザーが投稿した質問から機械学習を用いてユーザーの出産予定日を予測するシステムの構築に取り組んだ．

なお，研究の実施にあたり 5.3 節の実験結果では具体的な質問の分類結果について考察を行っているが，プライバシー配慮のために個人を特定できないよう，本論文で紹介する質問本文については一部文章の修正を行っている．

2. 関連研究

CQA サイトにおいて質問文から出産予定日を予測するという取り組みは行われていない．しかしながら，CQA サイトにおける質問分類の研究は行われてきた．

Kim ら[3]は Yahoo!Answers において，465 件の質問とその質問者からベストアンサーを与えられた回答について手作業で分類を行い，ベストアンサーの選択理由について考察を行っている．Kim らはそれらの質問を information（特定の事実の探索や現象の理解），suggestion（助言，推薦，実行可能な解決法の探索），opinion（他人の意見・感じ方の調査，議論の開始），others（先の 3 つのタイプに入らないもの）に分類した結果，それらの割合はそれぞれ information 35%， suggestion 23%，opinion 39%となった．また，ベストアンサーの選択理由の分布は質問のタイプによって違いがあることを示し，opinion タイプの質問に対しては，回答者の態度や感情面での支援など社会的・感情的な理由である socio-emotional の要素がベストアンサーの選択理由として大きな要素であるとしている．

栗山ら[4]は Yahoo!知恵袋において，500 件の質問を対象にして，手作業で質問のタイプを次の 3 つに分けた．1 つ目はサーチエンジンや図書館のレファレンス・サービスを利用して解答を探すことが可能な，「情報検索型」質問で，事実，真偽，定義・記述，方法・手段，原因・理由，効果・結果を尋ねる質問である．2 つ目は客観的な正解はなく，特定の個人あるいは集団に対してアンケート調査を行うことで回答を得るような「社会調査型」の質問で，助言，意見，施行，推薦，経験を尋ねる質問である．3 つ目は情報検索やアンケート調査によって客観的あるいは主観的な回答を得ることが目的ではなく，質問者が自分の主張に対する反響・反応を求めている記述表現を持つ「非質問型」で，主張もしくは記述としては何が書かれているのか分析者には理解できなかったものである．栗山らはユーザーの質問投稿時において，質問のトピックによるカテゴリ選択の支援だけでなく，これらの質問のタイプを提示することは利用者の支援になるという考えを示した．

CQA サイトにおいて，質問や回答のテキストデータを用いて特徴量を生成し，自然言語処理と機械学習を用いて，質問の分類を行っている研究には，Qu ら[5]や Aikawa ら[6]の研究がある．

Qu らは質問と回答のテキストデータを用いて, Yahoo!Answers における質問を SVM, Naive Bayes, Maximum Entropy を用いてカテゴリ分類を行っている．Aikawa らは Yahoo!知恵袋の質問において，質問テキストから取得できる特徴量のみで subjective な回答を求める質問か objective な回答を求める質問かの分類を，SVM と Naive Bayes を用いて行った．

質問や回答のテキストデータから得られる特徴量に加えて，Zhou ら[7] は like（ユーザーが有益と感じたときに回答に対してされるアクション），vote（ユーザーが最も良いとされる回答に対して投票を行なうアクション），source（回答の参照元），poll and survey（投票機能を用いた質問），answer number（回答数）といった要素から，手作業を用いずにシステムを利用して subjective な質問かどうかの分類を行った．

我々[8]は以前，共感を求める質問とそうでない質問の分類を行った．実社会において利用ニーズが存在している，共感を求める質問という質問タイプを定義し，質問テキストから取得できる特徴量のみを用いた分類を行うことで，単に分類方法を提案・評価するだけではなく，実応用可能な分類方法の開発を行った．また，実際にシステムに導入した例として，CQA サイトにおいて投稿されるコンテンツが違反コンテンツかどうかを検閲するシステムを構築した[9].

しかしながら，前述の通り CQA サイトにおいて質問文から出産予定日を予測するという取り組み，および類似した取り組みは行われていない．そこで，ユーザーが投稿した質問から機械学習を用いてユーザーの出産予定日を予測するシステムの構築を目標とする本研究に取り組んだ．

3. 質問文の特徴分析

3.1. 目的

本章では，女性向け Q&A コミュニティにおける質問文において，どのような質問文であれば投稿者の正しい出産予定日を把握することができるのかの分析を行う．その上で，課題があればそれらを明らかにし，どのようなアルゴリズム・前処理等が有効かを検討することが本章の目的である．

3.2. 対象とするデータ

今回利用するデータはママリ内において自分の子どもが生まれた生年月日を入力したユーザーが，出産以前に実際に投稿した質問である．ユーザーの子どもの生年月日から妊娠経過週を割り出すことにより各質問投稿時の妊娠経過週がわかる．これを教師ラベルとする．対象ユーザー数は 1,081 人，対象ユーザーが妊娠したときから出産までに質問した対象質問数は 21,886 である．

3.3. 問題設定

3.3.1. 目標とする指標の定義

本研究ではユーザーの妊娠経過週とともに，時系列でどの程度正確なユーザーの妊娠経過週の予測ができるのかを指標とする．ここで言う妊娠経過週とは，ユーザーの子どもの出産日から 280 日を引いて起算日とした値である．この妊娠経過週を「出産日を起算日とした妊娠経過週」とする．また，実際の出産日を起算日とした値ではなく，最終月経の初日から数えて 280 日目である出産予定日から 280 日を引いて起算日とした妊娠経過週を「予定日を起算日とした妊娠経過週」とする．

従って出産日を起算日とした妊娠経過週が正しく予測できれば出産日も算出することが可能である．出産日を起算日とした妊娠経過週ごとに予測を行うのは，早い段階で出産日を起算日とした妊娠経過週を予測できた方がその分長い期間ユーザーにマッチしたコンテンツや広告等を提供できるためである．

3.3.2. 正期産を考慮したユーザーの出産日を起算日とした妊娠経過週予測

予定日を起算日とした妊娠経過週が 37 週 0 日～41 週 6 日の間にお産があることを「正期産」[10]と言う．この期間に生まれた新生児は体の各機能も十分に成熟しているので，母体の外での生活にもスムーズに適応することができる．また，予定日を起算日とした妊娠経過 40 週 0 日を出産予定日と呼び，多くの人は正期産を迎える．正期産を考慮し，予測結果週の以前 3 週間，以後 1 週間と 6 日を出産日を起算日とした妊娠経過週の予測結果として，正しい予測結果であるとする．具体的な例として，例えば質問投稿者の出産日を起算日とした妊娠経過週が 39 週目でユーザーが質問の投稿を行ったとする．この場合，正期産を考慮し出産日を起算日とした妊娠経過週が 36 週目から 40 週と 6 日の間の週数とこの質問から予測されれば，出産日を起算日とした妊娠経過週の予測結果としては正しい予測結果であるとする．

3.3.3. 正解率と対象ユーザー率の定義

3.2 節で述べたデータ内の全ユーザー数に対して，3.3.2 項で述べた正期産を考慮したユーザーの出産日を起算日とした妊娠経過週が正しく行われたユーザー数の割合を「正解率」と定義する．

また，同様に対象とするデータ内の全ユーザー数に対して，出産日を起算日とした妊娠経過週の予測を行ったユーザー数の割合を「対象ユーザー率」と定義する．対象ユーザー率が 100%とならない具体的な事例については 3.5 節で述べる．

3.4. 特徴の分析方法

特徴の分析方法として，まずは質問文を目視で確認し，特徴の分析を行う．その内，ユーザーの出産日予測に有益と思われる方法を仮説立てし，実際にその方法で予測が可能かどうかを確認する．

3.5. 分析結果

質問を分析すると，具体的に自身の妊娠週を記し，それに応じた内容を回答者に尋ねている質問が存在することがわかった．例えば以下のような質問である．

「私は現在 4w の初妊婦です．最近妊娠がわかったのですが，つわりがとてもひどいです．皆さんはどうやってつわりを乗り越えましたか？」

質問文内の「4w」とは妊娠を話題としたテキストコミュニケーションの中で，予定日を起算日とした妊娠経過週が 4 週であることを示すためにユーザーが使う表現である．よってこの例の場合は，このユーザーは妊娠 4 週であると判断をすることが可能である．このように自身の妊娠週が質問文に含まれている場合は，正しくユーザーの出産日を起算日とした妊娠経過週を予測できる可能性が高いと考えた．以上より<数値> [w|週]という組み合わせが質問文内にある場合に正規表現を用いてその数値を週数として抜き出し，抽出した週数を出産日を起算日とした妊娠経過週とした．

妊娠してから出産に至るまで，ユーザーは週数に応じた質問をコミュニティの中で尋ねる．よって，多くの場合出産までの間にユーザーは複数回質問を行う．ユーザー1 人に対しての出産日を起算日とした妊娠経過週の予測は，予測時点でそれまでに投稿された質問を用いることが可能である．本章では予測時点の対象質問のうち，最も新しい質問の予測結果をユーザーの出産日を起算日とした妊娠経過週として採用した．結果を図 2 に示す．なお，各プロット上部の数値は出産日を起算日とした妊娠経過週を表す．

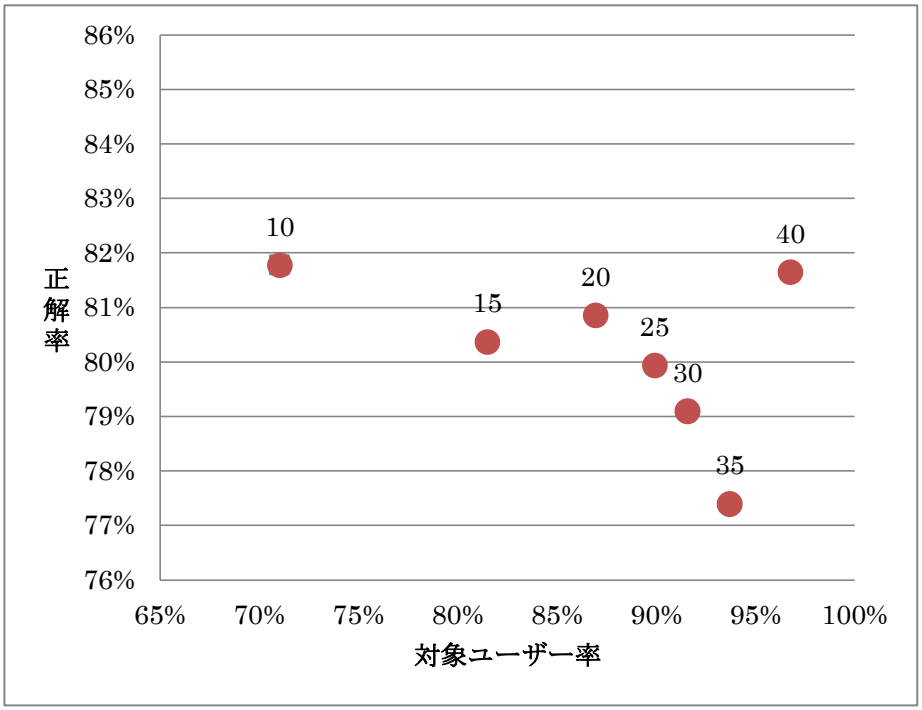

図 2 <数値> [w|週]という特徴語句を持つ質問から出産日を起算日とした妊娠経過週を抜き出し予測した正解率と対象ユーザー率

図 2 より，40 週目時点で対象ユーザー率は 95%以上となった．100%となっていないのは，ユーザーの中には<数値>[w|週]という組み合わせが質問文内に含まれた投稿を一度もせずに出産を迎えるユーザーも存在しているからである．

正解率について，図 2 を見ると正解率は 80%前後であった．これは自身の予定日を起算日とした妊娠経過

週に関係のない文脈で<数値>[w|週]という組み合わせが利用されているためだと考えられる．具体的には「今月 4 週目の土曜日に家族でディズニーランドに行く予定です！皆さんはもうハロウィンイベントに行かれましたか？」といった文章が質問文に存在する場合である．

そこで，正解率を上げるために，4w2d と言った形で<数値> [w|週]<数値>[d|日]という表現を持つ質問に絞って週数を抜き出した．「4w2d」とは，妊娠を話題としたテキストコミュニケーションの中で妊娠経過期間が 4 週 2 日であることを示すためにユーザーが使う表現である．

図 3 に抜き出した結果をまとめた．ここでは，各週数目における<数値>[w|週]と<数値>[w|週]<数値>[d|日]それぞれの組み合わせから週数を抜き出し予測した正解率と対象ユーザー率を示している．後者の結果から前者の結果へ矢印を引き，それぞれの週数における改善を示した．なお，図 2 同様各プロット上部の数値は出産日を起算日とした妊娠経過週を表す．正解率が上昇したことから，日単位までの情報を入れることで自身の状態について詳細に述べている質問文から週数を抽出することができたことがわかる．一方で週数の抽出条件を厳しくしたために 40 週時点での対象ユーザー率は 7 割程度となった．ここで対象ユーザー率よりも正解率を優先して考えることに関しては 4.1 節で解説を行う．

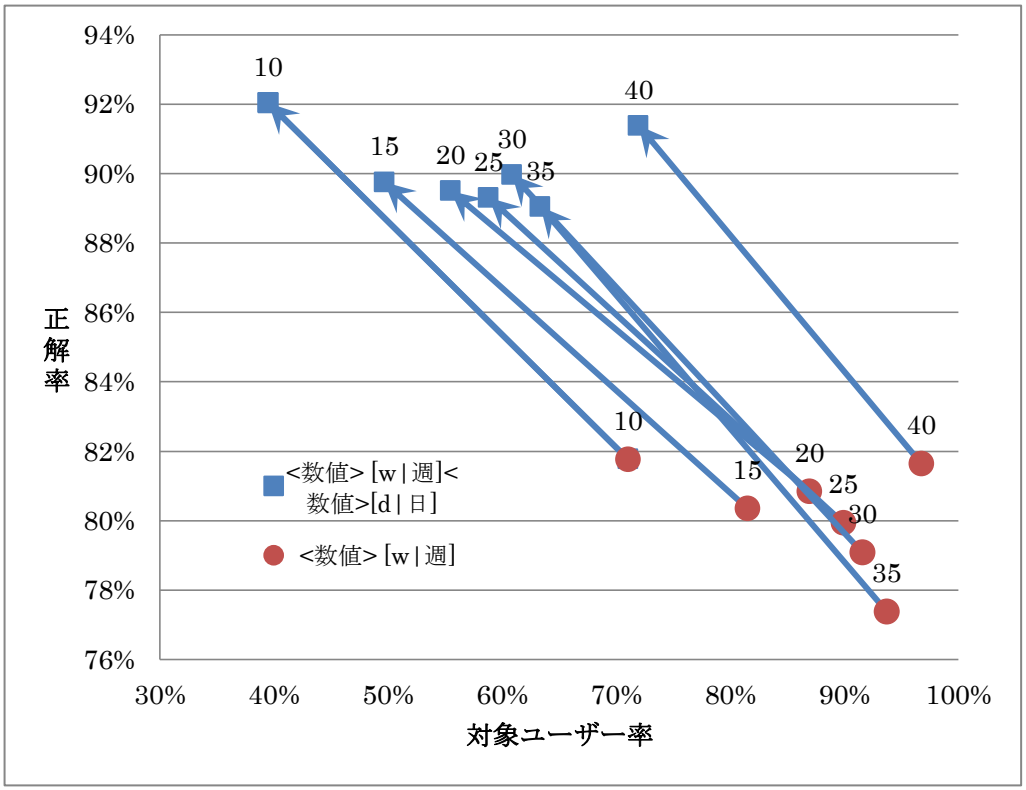

図 3 <数値>[w|週]と<数値>[w|週]<数値>[d|日]それぞれの特徴語句を持つ質問から出産日を起算日とした妊娠経過週を抜き出し予測した正解率と対象ユーザー率

4. 実システムへの導入のための判別手法の提案

4.1. 目標とすべき指標数値の定義

3.5 節より，正解率としては 90%を超える結果となった．しかし，それでも誤判別は約 1 割存在している．ママリアプリの利用者は約 80 万人以上存在している．従ってこのままの正解率でシステムを構築し，導入すると約 8 万人以上のユーザーには望まないコンテンツが表示され，ユーザーの不利益になる．そこでこの正解率を可能な限り 100%に近づけることを目標とすべき指標数値とする．

4.2. 提案手法の概要

4.1 節で定めた目標とすべき指標に近づくために機械学習を用いて，3.5 節で述べた，出産日を起算日とし

た妊娠経過週を，質問文から特徴語句を用いて抽出する前にフィルタリングを行うことにした．特徴語句を含んだ質問に対して，週数の推定として対象の質問が利用できるかどうかを機械学習によって分類する．提案手法のフローチャートを 図 4 に示す.

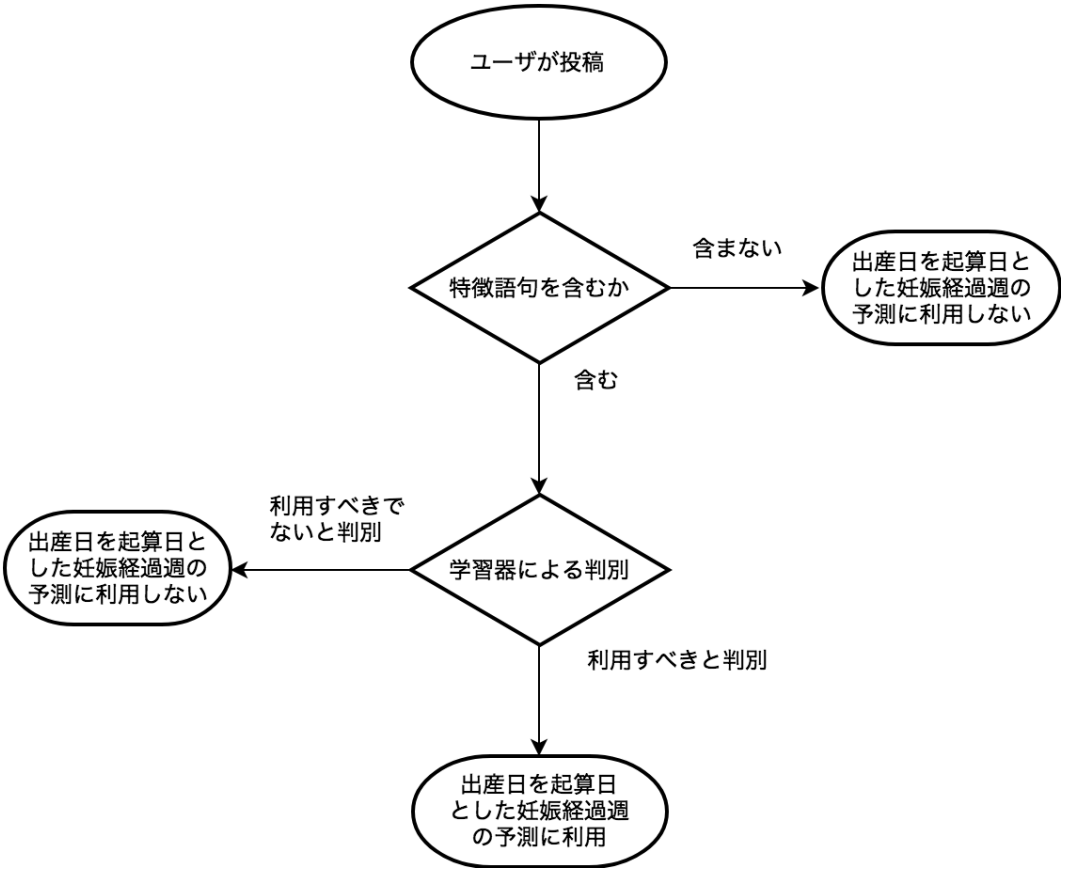

図 4 提案手法のフローチャート

4.3. 機械学習を用いた分類

3 章で述べた特徴語句による質問文からの出生日予測を行うべきかどうかを質問毎に機械学習を用いて分類する．自然言語処理における分類手法にはいくつかの手法が存在するが，

1. 形態素解析プログラム（MeCab[11]等）を用いて質問文から単語を抽出
2. 1 の単語から特徴量として用いる品詞として，名詞，動詞，形容詞，記号を抽出
3. 各単語を数値ベクトルへ変換する（これを単語ベクトルと呼ぶ）
4. 質問文毎に単語ベクトルの平均値を算出する
5. 4 を特徴量として学習器を用いて分類

という方法を用いて質問からの出生日予測を行うべきかを判別する．用いる単語の品詞としては，内容語である名詞，動詞，形容詞および，ママリ内では感情を表すものとしてよく使用されている記号を用いた．後者の例としては，「至急です！」中の「！」や「夜泣きで本当に悩んでいます。。。」の「。」が挙げられる．また，MeCab を用いて質問文から単語を切り出す際に利用される，システム辞書には mecab-ipadic-NEologd[12] を利用する．mecab-ipadic-NEologd とは，多数の Web 上の言語資源から得た新語を追加することにより拡張した MeCab 用のシステム辞書である. 各単語を単語ベクトルへ変換する方法として, word2vec[13], LSI[14], LDA[15]の内最も適したものを用いることとした．なお，LSI，LDA を使用する際には，単語文書行列には，単語の tf-idf 値を用いた．学習器には，linear，rbf カーネルを用いた SVM と Random Forest から最も適したも

のを用いることとした．手順 3 で利用する word2vec，LSI，LDA それぞれのモデル作成にはママリに投稿された約 100 万件の質問をコーパスとし, 学習データとして利用した (以下, ママリコーパスと呼ぶ)．word2vec のモデル作成において，特徴ベクトルの次元数は 100，文脈学習時の前後の対象単語の幅は 5 とし，ママリコーパスに出現する頻度が 5 回未満の単語は使用しなかった．また，LSI のモデル作成におけるトピック数は 200，LDA のモデル作成におけるトピック数は 100 である．実装には gensim[16]を利用し，前述のそれぞれのモデル作成時のパラメータはライブラリのデフォルト値である．学習器と単語ベクトルの生成方法の組み合わせによる実験で性能を比較し，最も良い結果を目指す.

5. 実験

5.1. データ・セット

対象データ・セットとしては 3.2 節で述べたデータと同じものを利用する．ママリ内において自分の子どもが生まれた生年月日を入力したユーザーが実際に投稿した質問を用いているので，それぞれの質問には実際の出産日を起算日とした妊娠経過週のラベルを持つ．データ・セットから 3.5 節の分析の結果発見した，<数値> [w|週]<数値> [d|日]という表現を持つ質問が，本実験で利用するデータとなる．<数値> [w|週]<数値> [d|日]という表現を持つ質問の数は 5,198 個．その内 3.3.2 項で述べた正期産を考慮した出産日を起算日とした妊娠経過週予測で正しく週数を抜き出した質問の数は 4,686 個, 誤って抜き出してしまった質問の数は 512 個である．質問文から抽出した実験に用いた名詞, 動詞, 形容詞, 記号の数はそれぞれ名詞 590,065 語, 動詞 297,273 語, 形容詞 43,979 語, 記号 244,440 語である．これらの質問に対して機械学習を用いて出産日を起算日とした妊娠経過週の予測に利用するかの分類を行う.

5.2. 実験手法

4 章で述べた提案手法を用いて 5.1 節のデータ・セットに対して機械学習を用いて出産日を起算日とした妊娠経過週の予測に利用するかの分類を行う．実装には scikit-learn [17] を利用し，学習器それぞれのパラメータについては，学習器と単語ベクトルの生成方法の組み合わせごとに学習データを用いてグリッドサーチを行い, 最適な値を求めた. なお, 以後の実験での評価は, すべて 10-fold cross validation で行った. また, scikit-learn の SVC，RandomForestClassifier を用いて probability を出力させ，probability がある値（以下，閾値）以上であれば出産日を起算日とした妊娠経過週の予測に利用する，そうでなければ利用しない．この閾値を調節することにより，出産日を起算日とした妊娠経過週予測の正解率を向上させることにした．probability を利用した理由は，同一人物が特定の出産日を起算日とした妊娠経過週において，複数個対象となる質問が見つかるかもしれないからである．その場合は特定の出産日を起算日とした妊娠経過週で，最も probability が高い質問を採用することで正解率を上げることができると考えた．なお，10-fold cross validation で probability の閾値を学習データから動的に出力するために，学習データの中で 10-fold cross validation を再度行い，正解率が 100%になった閾値の平均値を採用することにした．probability の閾値の導出方法を図 5 に示す.

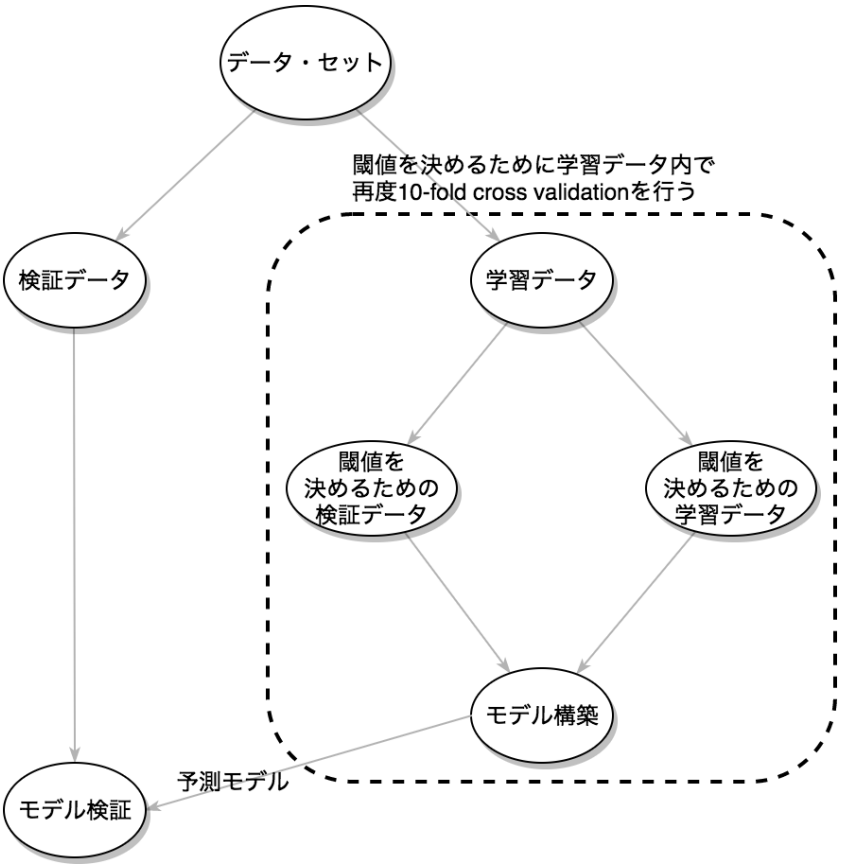

図 5　probability の閾値の導出方法

5.3. 実験結果

図 6 に学習器と単語ベクトルの生成方法の組み合わせによる，出産日を起算日とした妊娠経過週が 40 週時点の正解率と対象ユーザー率の比較を示す．図中で組み合わせを示すために RandomForest は RF，word2vec は W という略称を用いた．グリッドサーチによるパラメータチューニングの結果，学習器と単語ベクトルの生成方法の組み合わせによるそれぞれのパラメータは SVM(rbf)-LDA: C=1000，gamma=0.001，SVM(rbf)-LSI: C=10，gamma=0.001，SVM(rbf)-W: C=1000，gamma=0.0001，SVM(linear)-LDA: C=1，SVM(linear)-LSI: C=1，SVM(linear)-W: C=10，RF-LDA: min_samples_leaf=200，n_estimators=2000，RF-LSI: min_samples_leaf=200，n_estimators=1000，RF-W: min_samples_leaf=1000，n_estimators=3000 となった．図 6 で最も正解率の高かった RF-W という組み合わせの実験結果を提案手法とし，<数値>[w|週]<数値>[d|日]の特徴語句を持つ質問から出産日を起算日とした妊娠経過週を抜き出し予測した正解率と対象ユーザー率の比較を図 7 に示す．

図 7 において矢印の始点となっている■の点は，3.5 節の<数値>[w|週]<数値>[d|日]という表現を持つ質問に絞って週数を抜き出した実験結果である．▲が提案手法の結果となっており，それぞれに向かって矢印を描いている．提案手法は 40 週時点では正解率 95%を超える結果となった．また，妊娠 40 週時点での対象ユーザー率は約 20%となった．これは 4.1 節で述べた通り，望まないコンテンツのレコメンデーションはユーザーにとって不利益となるため，対象ユーザー率よりも正解率を優先し，モデル作成の中で正解率が 100%になった閾値を採用した結果である．

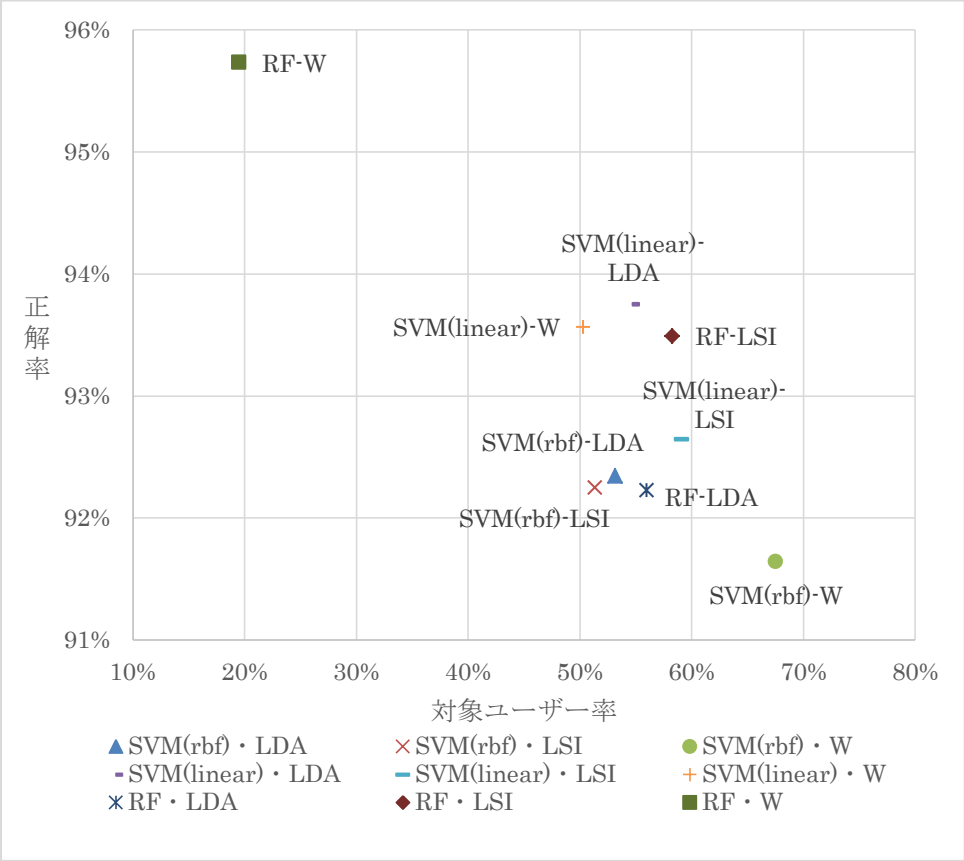

図 6　学習器と単語ベクトルの生成方法の組み合わせによる，出産日を起算日とした妊娠経過週が 40 週時点の正解率と対象ユーザー率の比較

提案手法により正しく妊娠週数を抽出できている質問を見ると，正期産でなかった兆候を質問文から判別できていたことが分かった．具体的には以下のような質問である

「33w2d です．前回の検診では大きな問題はなかったですが、先生からコメントいただいた点で気になったのは子宮頸管の長さが正常の中でも短い方と言われました．同じように言われた方はいらっしゃいますか？少し歩くだけでお腹が張ってしまうので、今は安静にしています．次の検診は明日です．今少し張りが強くて，病院に電話しようか迷っています。皆さんはどんな症状で病院に電話しましたか？もしできれば、同じような経験をされた方からコメントもらえると嬉しいです.」

このような質問を投稿したユーザーはこの投稿後の 2 週間後（35 週目）で出産を迎えたので，早産の兆候を含む質問を的確に抽出できた例だと考えられる．一方で提案手法でも誤判定をしてしまった質問を見ると，正期産でなく早産であったり，妊娠満 42 週以降の出産であったりするユーザーの質問であり，その兆候は人が見ても質問文から判別しづらい質問であることがわかった．具体的には以下のような質問である．

「24w5d です！最近食欲がやばいです！とくに今週…!!! 元々スレンダーではない体系なのですが，頑張って，かなりいろいろ大変だったのですが，なんとか体重の増加は 1.4 程度におさまっています＼(^o^)／でも出産までまだ数ヶ月あるのが恐怖でなりません。･ﾟ･(ﾉД`)･ﾟ･。皆さんどうやって体重コントロールされましたか？」

このような投稿を行ったユーザーの中には 44 週目で出産を迎えているユーザーが存在している．正期産でないが，文章からその兆候が見られない．このような質問文の存在が，妊娠経過週によって 100%を達成できなかった理由であると考えられる．

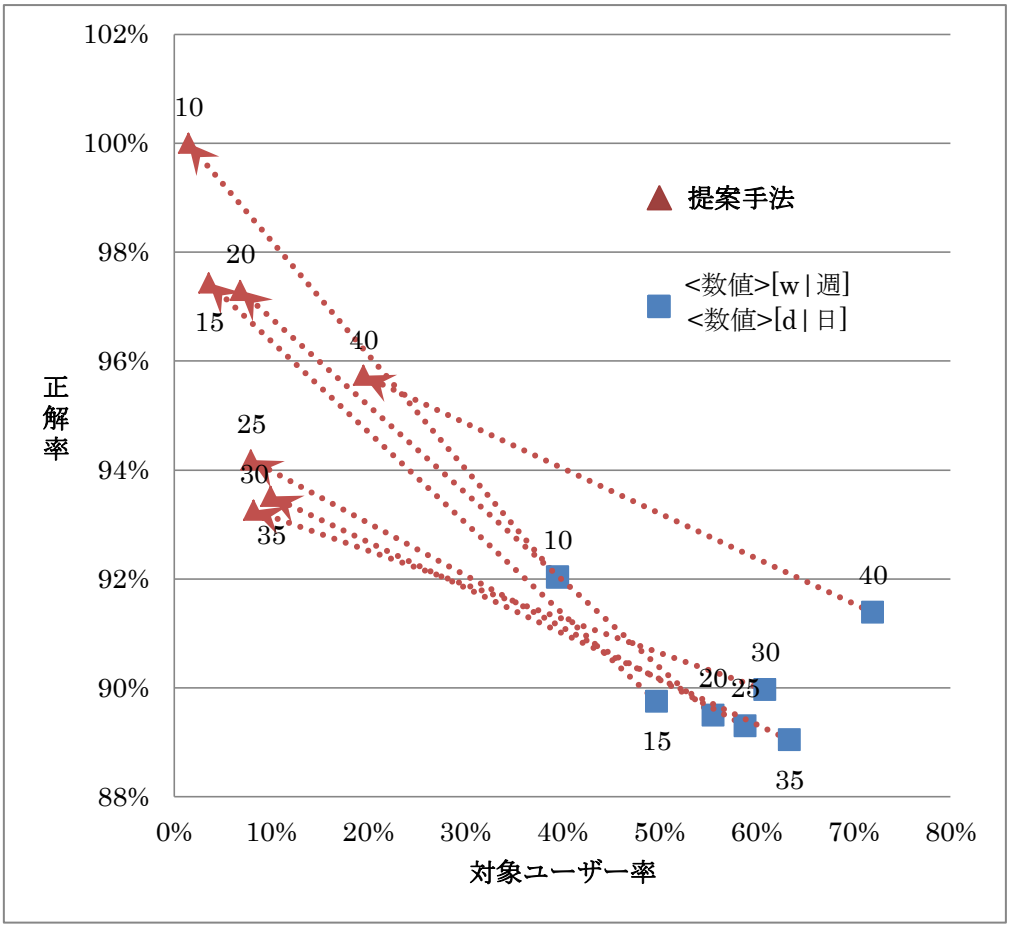

図 7 提案手法と<数値>[w|週]<数値>[d|日]の特徴語句を持つ質問から出産日を起算日とした妊娠経過週を抜き出し予測した正解率と対象ユーザー率の比較

6. 終わりに

本研究では CQA サイトにおいて出産日を入力せずとも，ユーザーが投稿した質問からユーザーの出産予定日を予測するシステムの構築を目的として，質問文の特徴分析や実システム導入のための判別手法を提案した．質問文の特徴分析では出産日の予測に役立つ要素表現を見つけ，検証を行った．さらに正解率を高めるために，出産予定日予測のための出産日を起算日とした妊娠経過週抽出前に機械学習を用いた事前の分類フィルタを出産予定日予測の処理フローに組み込んだ．その結果，95%を超える正解率で出産予定日の予測を行うことができた．今後は実際にプロダクトに組み込み，ユーザーへの影響を見ながら，ユーザーの満足度を向上させることや，収益化に取り組んでいきたいと考えている．

参考文献

[1] ママリ : https://qa.mamari.jp/, 2017.10.19 参照.
[2] 厚生労働省 平成 27 年人口動態統計月報年計（概数）の概況 : http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/geppo/nengai15/index.html, 2017.10.19 参照.
[3] Kim, S., Oh, J. and Oh, S., “Best-Answer Selection Criteria in a Social Q&A site from the User-Oriented Relevance Perspective,” American Society for Information Science and Techonology (ASIS&T) 2007 Annual Meeting, Milwaukee, Wisconsin, ASIS&T, 2007.
[4] 栗山和子, 神門典子, “Q&A サイトにおける質問と回答の分析,” 情報処理学会研究報告, Vol. 2009-FI-95, No.19, 2009.
[5] Bo Qu, Gao Cong, Cuiping Li, Aixin Sun, and Hong Chen, “An evaluation of classification models for question topic categorization,” JASIST, Vol.63, No.5, 2012, pp. 889-903.
[6] Naoyoshi Aikawa, Tetsuya Sakai, and Hayato Yamana, “Community qa question classification: Is the asker looking

for subjective answers or not? ,” IPSJ Online Transactions, Vol.4, 2011, pp. 160-168.
[7] Tom Chao Zhou, Xiance Si, Edward Y. Chang, Irwin King and Michael R. Lyu, “A data driven approach to question subjectivity identification in community question answering,” in Proc. AAAI, 2012.
[8] 島田達朗, 櫻井彰人, “日本最大級の女性向けコミュニティサイトにおける違反コンテンツ検閲システムの構築,” 情報処理学会デジタルプラクティス, Vol.8, No.1, 2017, pp. 92-96.
[9] 島田達朗, 櫻井彰人, “コミュニティサイトにおける共感を求める質問の認識,” 知能と情報, Vol.29, No.4, 2017, pp.611-618.
[10] 岡井崇, 綾部琢哉（編）, “標準産科婦人科学,” 医学書院, 2014.
[11] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), 2004, pp. 230-237.
[12] Sato, T., Neologism Dictionary Based on the Language Resources on the Web for Mecab : https://github.com/neologd/mecab-ipadic-neologd, 2017.10.19 参照.
[13] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, “Efficient estimation of word representations in vector space,” ICLR Workshop, 2013.
[14] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, “Latent semantic indexing: A probabilistic analysis,” Journal of Computer and System Sciences, Vol.61, No.2, 2000, pp. 217-235.
[15] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research Vol.3, 2003, pp. 993-1022.
[16] gensim : https://radimrehurek.com/gensim/index.html, 2017.10.19 参照.
[17] scikit-learn : http://scikit-learn.org/, 2017.10.19 参照.

著者略歴

島田 達朗（しまだ たつろう）

2011 年慶應義塾大学理工学部管理工学科卒業.

2013 年同大学理工学研究科前期博士課程修了．在学中に Connehito 株式会社を創業，取締役．現在に至る.

櫻井 彰人（さくらい あきと）

1975 年東京大学工学部計数工学科卒業．1977 年同大学大学院情報工学研究科修了.

1977 年株式会社日立製作所入社那珂工場配属．1989 年同基礎研究所，1996 年同中央研究所，1998 年北陸先端科学技術大学院大学教授を経て，2001 年慶應義塾大学理工学部教授となる.

2018 年 3 月をもって定年退職した．博士（工学），慶應義大学名誉教授.