

混合ユニグラムモデルによる質問紙調査データ分析

鳥取大学大学院 学生会員 ○嶋津 裕樹 鳥取大学大学院 正会員 谷本 圭志
鳥取大学大学院 正会員 長曾我部 まどか 鳥取大学大学院 正会員 土屋 哲

1. はじめに

質問紙調査データの分析では、分析者の仮説に基づき、少数の質問項目が抽出され、回答者の属性と回答内容の違いといった項目間の関係が解析される場合が少なくない。この場合、分析者が限定した項目のみが扱われることになり、分析者の仮説にない潜在的な項目間の関係については見落とされる懸念がある。近年、文書解析の手法として、潜在意味解析^{1),2)}が提案されている。これは、文書データが有する意味を統計的に解析する方法であり、自然言語処理の分野で開発された。近年では、文書データに限らず、購買データにおけるユーザーの嗜好なども研究の対象となっている。畠山ら³⁾は、健康診断における問診項目情報に対して、潜在トピックモデルを適用している。本研究では、統計的潜在意味解析手法の一つである混合ユニグラムモデルを質問紙調査データに適用し、項目間の潜在的な関係を明らかにすることを目的とする。

2. 本研究の基本的な考え方

潜在意味解析では、複数の単語の共起性によって推計される情報を「潜在の意味」と考え、潜在の意味のカテゴリを「潜在トピック」と呼ぶ。トピックごとに異なる単語分布 ϕ があるとし、文書の集合 D からトピック z とトピックに含まれる単語 w を推計する。本研究では、混合ユニグラムモデルを使用する。図1にモデルの概要を示す。灰色の円は観測変数、白地の円は未知変数を示す²⁾。四角は繰り返しを表し、右下の記号はその数を表す。ここで、単語 w は観測変数であり、トピック z 、トピック分布 θ 、単語分布 ϕ 、ハイパーパラメータ α, β は未知数である。右側の四角はトピック毎の単語分布 ϕ がトピック数 K あることを表す。また左外側の四角は D 文書あること、左内側の四角はそれぞれの文書に N 単語含まれることを表す。混合ユニグラムモデルでは、文書ごとに1つのトピック z があることになる。

本モデルを質問紙調査に適用する場合、質問項目に

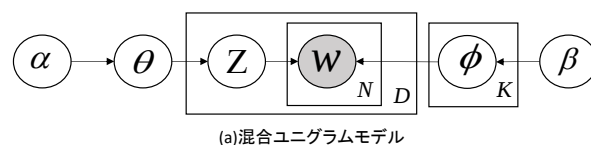


図1 混合ユニグラムモデル¹⁾

対する、複数の選択肢の共起性によって推計されるトピックを「潜在クラス」と考える。その上で、各クラスにおける特性(クラス、回答の傾向)を明らかにする。

3. 混合ユニグラムモデル

本研究では、質問紙の回答者を d とし、質問項目についてのクラス k に関する選択肢分布を $\phi_k=(\phi_{k1},\dots,\phi_{kv})$ とする。選択肢とは、性別を質問した場合の「男性」や「女性」を示す。ここで、 $\phi_{kv}=p(v|\phi_k)$ は、クラス k において、選択肢 v が出現する確率である。 $\theta_k=p(k|\theta)$ は、回答者にクラス k が割り当てられる確率である。混合ユニグラムモデルでは、クラスごとに異なる選択肢分布 ϕ_k を持つため、それぞれのクラスで出やすい選択肢を表現できる。パラメータ θ, Φ が与えられたときの回答者 d が w_d を回答する確率は以下のように定式化される。

$$p(w_d|\theta, \Phi) = \sum_{k=1}^K \theta_k \prod_{v=1}^V \phi_{kv}^{N_{dv}} \quad (1)$$

ここで、 N_{dv} は回答者 d について、選択肢 v が現れる回数を表す。本研究では混合モデルのパラメータを最尤推定する方法としてEMアルゴリズムを用いる。EMアルゴリズムは対数尤度の下限を最大化することでパラメータの局所最適値を求めるものである。混合ユニグラムモデルの対数尤度を次式に示す。

$$L = \sum_{d=1}^D \ln p(w_d|\theta, \Phi) = \sum_{d=1}^D \ln \sum_{k=1}^K p(k|\theta) p(w_d|\phi_k) \quad (2)$$

キーワード 混合ユニグラムモデル, 質問紙調査

連絡先 〒680-8552 鳥取市湖山町南4-101 鳥取大学大学院 持続性社会創成科学研究科工学専攻

T E L 0857-31-5311

他に、情報量基準の適合度の指標として AIC, BIC を用い、クラス数を決定する。

4. 事例分析

(1) データの概要 2016年10月に鳥取県A町で実施した地域運営組織に関する質問紙調査の結果に対し、モデルを適用する。質問は25項目、選択肢の合計は196である。回答者数はN=153である。質問紙では、回答者の属性、回答者の平日と休日の過ごし方、組織が運営する施設の認知度と利用状況などを尋ねた。質問内容の概要を表1に示す。

(2) 分析結果 回答者は3つのクラスに分類された。回答者の属性に関する結果を図2に示す。クラス1は「男性」、「60代・70代」、「無職・年金生活者」という特性を示した。また「平日と休日の外出が多い」という結果も得られた。クラス2は、「女性」、「60代・70代」、「無職・年金生活者」という特性を示した。また「外出が少ない」という結果も得られた。クラス3は、「40代・50代」、「会社員・団体職員」という特性を示した。また、組織が運営する施設の認知度と利用の頻度についての結果を図3に示す。図3より、クラス3は「施設を定期的に利用している」、クラス1は「施設を利用していない」という特性が明らかになった。すなわち、普段から外出が多い高齢者は施設を利用し、普段の外出が少ない高齢者は施設を利用しない、といった特性が示唆された。

また、同データに対し、k-means法による非階層型クラスター分析を適用した結果、類似した結果が得られたことから、混合ユニグラムモデルによる質問紙調査データ分析は有用であると考えられる。また、クラスター分析と違い、計算により最適なクラス数を求めている点において本研究のアプローチが優れていると考えられる。

参考文献

- 1) 奥村学（監修），トピックモデルによる統計的潜在意味解析，コロナ社，2015.
- 2) 岩田具治，トピックモデル，講談社，2015. 3
- 3) 畠山豊，他，問診データに対する潜在トピックモデルに基づく健診データ解析，医療情報学，33(5):267-277, 2013.

表1 質問の概要

① 回答者の属性について
・性別，年齢，住まいの集落，職業，家族構成，普段の移動手段
② 平日と休日の過ごし方について
・平日，休日の外出頻度，外出している時間帯，外出理由
・集落の共同作業や催しの参加有無
・地区での生活満足度
③ 施設との関わりについて
・施設について知っているか
・施設が行っているサービス，活動について知っているか
・施設を知ったきっかけ
・施設の利用頻度（頻度により「利用者」と「非利用者」に分類）
<利用者への質問>
・月あたりの利用頻度
・利用目的
・施設のイベントで，参加したことがあるものについて
<非利用者への質問>
・利用していない，したことがない理由について
・どのような活動やサービスがあれば利用したいか

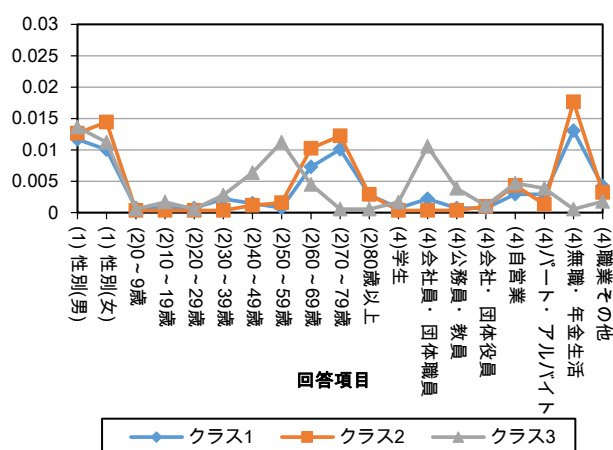


図2 回答者の性別・年齢・職業

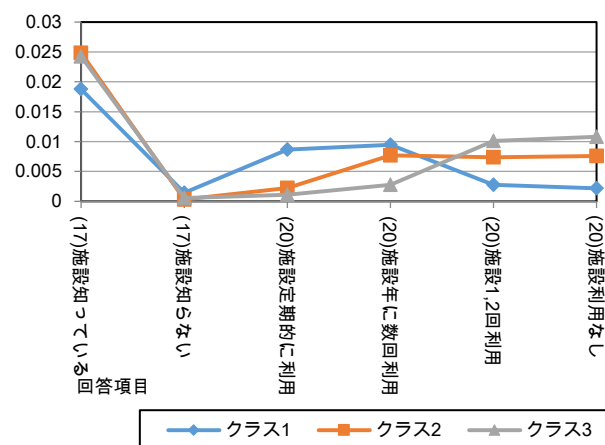


図3 回答者の施設利用の有無