質問紙調査データへの混合ユニグラムモデルの適用可能性に関する一考察

公共システム研究室　嶋津裕樹

1. はじめに

文書解析の手法として，潜在意味解析が提案されている．これは，文書データに潜む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．そこで本研究は，従来の手法で課題とされてきた点を解決する方法として，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，その分析可能性を検討する．

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「潜在的意味」と考える．混合ユニグラムモデルでは，トピックごとに異なった単語分布を持っているとし，文書集合からトピックとトピックに含まれる単語を推計する．本研究では，質問紙調査における質問項目に対する，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. 分析手法

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 $\phi_k=(\phi_{k1},\cdots,\phi_{kV})$ とする．選択肢とは，性別を質問した場合の「男性」や「女性」のことである．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を用いると，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．

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

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．

4. 事例分析

2016 年 10 月に大山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 個である．質問紙の回答者数は，N=153 であった．

分析の結果，3 つのクラスに分類された．回答者の基本的な情報を図 1 に示す．クラス 1 は「男性」，「60 代・70 代」，「無職・年金生活者」で，「平日と休日の外出が多い」ことが明らかになった．クラス 2 は，「女性」，「60 代・70 代」，「無職・年金生活者」で，「外出が少ない」ことが明らかになった．クラス 3 は，「40 代・50 代」，「会社員」，「核家族世帯」，であることが明らかになった．また，組織が運営する施設の認知度と利用頻度についての結果を図 2 に示す．図 2 より，クラス 3 は「施設を定期的に利用している」こと，クラス 1 は「施設を利用していない」ことが明らかになった．

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，計算結果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．

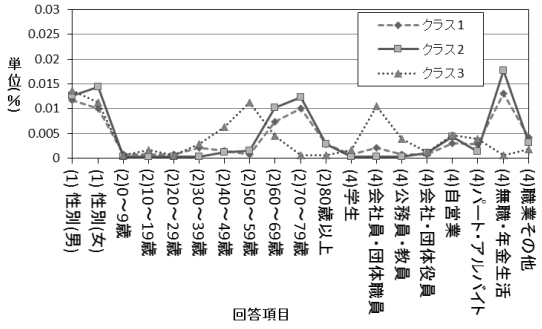

図 1　回答者の性別・年齢・職業

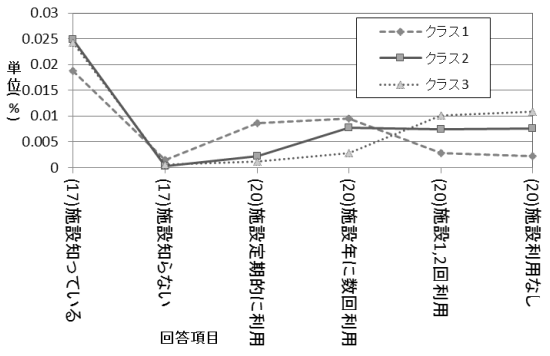

図 2　回答者の施設利用の有無