

データベースを作成して雑誌分析をする際の手法の比較

——テキストマイニングと量的内容分析——

東京大学大学院 塩谷昌之

【1. 目的】

本報告の目的は、同一の雑誌資料を分析するに当たって、テキストマイニングと量的内容分析という二つの手法を比較することで、それぞれの利点を具体的な分析結果と共に反省的に示すことである。報告者はかつて少年向け科学雑誌である『子供の科学』（誠文堂新光社）の目次欄に準拠し、1924～60年の記事タイトル15250件を網羅的にまとめたデータベースを作成した。実際にこれに二つの手法それぞれを適用する中で得られた論点を、特に作業上のコストとその成果との関わりに注目しながら報告する。

【2. 方法】

テキストマイニングは機械的処理による言葉のカウント、量的内容分析は人為的処理による内容のカウントを行うものである。本報告では上記のデータベースを、それぞれの手法に適した形へと加工した上で、分析精度、分析上の利点、作業コストなどの比較を行う。二つの手法には分析上のコードを後で与えるか、先に与えるかの相違があるため、実際の調査ではまずテキストマイニングによって得られた頻出語の分布を参考に、変数の形式やラベルといった分析コードを検討し、続く量的内容分析でより精密なコーディングを施すという順番を辿った。テキストマイニングの具体的な手続きについては報告時までには複数のものを再検討する。目次に準拠するため、記事タイトルと本文記事との不一致や、内容の齟齬については照合していない。

【3. 結果】

テキストマイニングは資料の全体的特性を明らかにすること、残余カテゴリーから意識外の項目を拾い上げることには有効であった。記事タイトルが短文であったため共起性の検討はできなかったが、辞書を設定することで複数の言葉を集約し、分布の変化を明らかにすることも可能であった。一方で、量的内容分析もまた集計結果から分布の変化を示すことが可能であったが、さらに複数の変数を設定して、それらの関係性を丁寧に見ることに適していた。手法ごとの分析精度については結果にあまり差異が見られず、これは人為的処理においても記事タイトルの単語を参考にコードを振り分けていたためだと考えられる。注目すべきは原資料のデザインと、そこから読み込める情報の階層性である。実際にデータ化を行った際にはテキストだけでなく目次欄のデザイン、すなわち枠線の有無、インデント、フォントの大小などによって、特定の記事がどの項目に振り分けられていたのかを解釈する必要に迫られた。テキストの水準のみでこれに対応するのは現段階では困難であり、量的内容分析の方がデザインを含めた原資料の意図を再現するのに適していると言える。

【4. 結論】

資料の特性や分布の変化をマッピングするにはテキストマイニングを、原資料のデザインの含意を再現した上で変数間関係性を見るには量的内容分析を、より精緻な分析を行うためには双方を両立させるのが望ましい。しかし、いずれもデータ化の段階から考えるとかなりの作業上のコストが必要になる。目的に応じた手法に絞ることで、コストを抑えた分析が可能になるだろう。

文献

神野由紀・辻泉・飯田豊編、2019、『趣味とジェンダー——<手づくり>と<自作>の近代』青弓社。