社会科学文献のテキストマイニング

―テキストマイニングの社会学的利用 1―

東北学院大学　鈴木努

1　目的

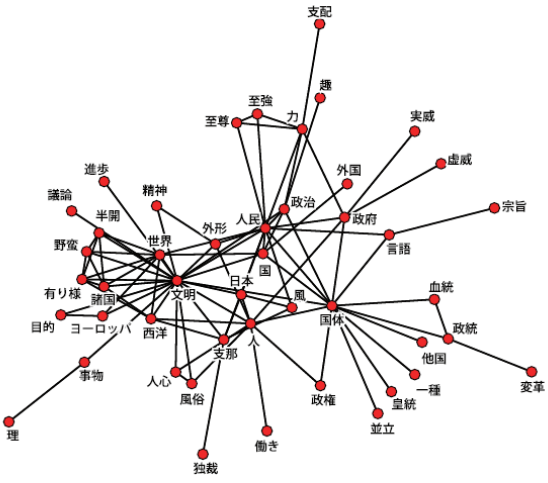
テキストマイニングは文書データから有用な情報を発見する方法として社会学的な分析にも広く用いられるようになっている．従来から新聞記事や自由記述回答の分析にはこのような方法が用いられてきたが，最近ではウェブ上に蓄積されたビッグデータやデジタルアーカイブを対象とした分析が注目されている．例えば，Google Books にアーカイブされた古今の書籍に現れる単語の頻度を調べることで，ある概念の登場が歴史学における通説とは異なっていたことが示されたり（Aiden & Michel 2013），デジタルデータ化された米大統領の一般教書演説を分析することで主要政策の変遷が可視化されたりしている（Bearman 2015）．一方，英語のように単語が分ち書きされない日本語のテキストでは，単語への分割という作業が必要なことに加え，歴史的資料においては漢字の旧字体，古典文法への対応などテキストマイニングするうえで困難な点が多い．しかし，デジタルアーカイブや分析用辞書の整備は進められており，人文社会科学における歴史的資料に対するテキストマイニングの適用は今後さらに進んでいくと考えられる．本報告では近代文語で書かれた文献にテキストマイニングを適用した例を示し，従来の読解と比較することでその意義と可能性を示す．

2　方法

分析例としてテキストデータの入手の容易さ，分析結果と比較できる注釈書の存在から，福沢諭吉『文明論之概略』を用いる．ただし，丸山真男『「文明論之概略」を読む 上』の第 3 講から第 5 講で論じられている『文明論之概略』第 2 章に限って分析する．テキストデータは上田修一・慶應義塾大学教授のウェブサイトで公開されていたものを使用した．形態素解析には「近代茶まめ」を用い，生起頻度や共起頻度の分析，共起ネットワークの可視化には R を用いた．

3　結果

丸山は第 2 章について 3 つの講を充て，それぞれ「文明」「自由」「国体」に焦点を当てて論じている．しかし，テキストマイニングにより単語（名詞）の共起関係をネットワークとして表すと（右図），「文明」と「国体」は中心的な概念としての位置にあるものの，「自由」はそれらと同等の位置を占めていない．このことから「自由」を重視する視点は丸山の読解の特徴であると考えられる．このように社会科学の文献にテキストマイニングを適用することで，学史や学説の研究に新しい視点を提供すること期待される．



文献

Aiden E. & Michel J., 2013, *Uncharted*.（＝2016，阪本芳久訳『カルチャロミクス』草思社.）
Bearman, P., 2015, “Big Data and Historical Social Science,” *Big Data & Society*, 2(2): 1-5.