2G-03

計量テキスト分析を用いたQ&Aコミュニティからの評判情報の推定

吉見憲二†
佛教大学†

1. はじめに

Q&A コミュニティはソーシャルメディア等のオンラインコミュニティと比較して，より相談者が具体的な目的意識を持って投稿していることが予想される．そのため，単純な投稿数をベースとした情報量ではソーシャルメディア等に劣るものの，特定の分野では企業等に対してより重要な情報を提供できる可能性がある．本研究では，そのような問題意識からQ&Aコミュニティの中でも多くのユーザに利用されている Yahoo!知恵袋の投稿データを用いて，計量テキスト分析の手法から評判情報を推定することを試みる．

2. 分析内容

2.1. 分析対象

本研究では，分析対象として国立情報学研究所が提供する Yahoo!知恵袋の質問者データを用いた[1]．Yahoo!知恵袋は日本最大の Q&A コミュニティであり，多くのユーザーから質問と回答が寄せられている．大規模データ分析にあたっては Twitter や Facebook 等のソーシャルメディアの投稿者データが用いられることが多いが，Q&A コミュニティの投稿者データではそれらに比べてより具体的な問題意識に沿ったものを入手することが期待できる．今回は特に投稿者の目的がはっきりしている「大学受験」のカテゴリを対象とした．

2.2. 分析目的

「大学受験」カテゴリでは多くの大学名が言及されており，受験校の選定に悩んでいる受験生やその保護者から複数の大学名が比較の文脈で登場することも少なくない．そこで，本研究では大学名の共起関係から投稿者の大学に関する主観的な評価を可視化することを試みた．ただし，全国の大学名を扱うことは難しいため，一般によく知られている大学群の名称を中心に検討する．

具体的には，関東の私大のグループである「早慶上(理)」「MARCH」，及び，関西の私大のグループである「関関同立」「産近甲龍」及びそれらのグループに含まれる大学名を対象とする．

2.3. 分析手法

分析に当たってはフリーのテキストマイニングソフトウェアである KH Coder（https://khcoder.net/）を使用した．形態素解析は付属の ChaSen（茶筌）を用いているが，今回は強制抽出の対象とした語句のみを分析している．抽出対象の語句は基本的に「○○大学」の○○の部分としているが，「関西大学」と「近畿大学」はそれぞれ地名を含むことから「関西大」「近畿大」を抽出対象とした．「青山学院大学」と「関西学院大学」は略称である「青学」「関学」も抽出対象としている．

分析の手順として，まず対象語句の出現数を算出し，次に抽出した語句の登場傾向を「コレスポンデンス分析（対応分析）」で可視化した．最後に，単語間の共起関係について共起ネットワーク分析を用いて可視化した．共起ネットワーク分析の描写に当たっては Jaccard 係数 0.1 以上の共起関係を基準とした．

2.4. 仮説

本研究の仮説として，以下の 2 点を提示する．
仮説 1：大学名の共起関係は大学群のカテゴリに影響されている．
仮説 2：関東と関西で大学名の共起関係の傾向に差異がある．

3. 分析結果

3.1. 大学名の出現数

提供データにおける「大学受験」カテゴリの投稿数は 35,597 件であった．各大学名の出現数は表 1 の通りである．一番言及数の多かった「早稲田」で全投稿に占める比率は 4%程度となった．なお，ここでの出現数は投稿中の出現の有無のみをカウントしているため，1 投稿に複数回登場しているものであっても 1 回となる．

Estimating reputation information from Q&A communities using quantitative text analysis
†Kenji YOSHIMI, Bukkyo University

表１　大学名の出現数

大学名	出現数	大学名	出現数
早稲田	1435	関西大	489
MARCH	934	立教	449
明治	916	近畿大	403
同志社	846	青山学院	397
慶應	828	関学	338
立命館	705	龍谷	272
法政	650	東京理科	245
関西学院	648	甲南	223
早慶	523	京都産業	182
中央	506	青学	175
上智	503	産近甲龍	140
関関同立	495	早慶上智	113

3.2. コレスポンデンス分析

コレスポンデンス分析の結果は図１の通りとなった．概ね「早慶のグループ（右下）」「MARCH＋東京理科大学のグループ（右上）」「関西の大学のグループ（左中央）」という傾向が見られている．横軸を構成する成分１は関東と関西の差異と解釈でき，縦軸を構成する成分２はいわゆる偏差値に近いものであると解釈できる．ただし，関西の大学群においては成分２による差異はあまり大きく見られなかった．

コレスポンデンス分析では大学群によるグループ分けが概ね支持される結果となっている．しかしながら，成分 1 の寄与率が 7.31%，成分 2 の寄与率が 6.37%であり，累積寄与率が 13.68%に留まる点には注意が必要である．これは全体の投稿に占める対象となる大学名を含む投稿が必ずしも多くないことが影響しているものと考えられる．

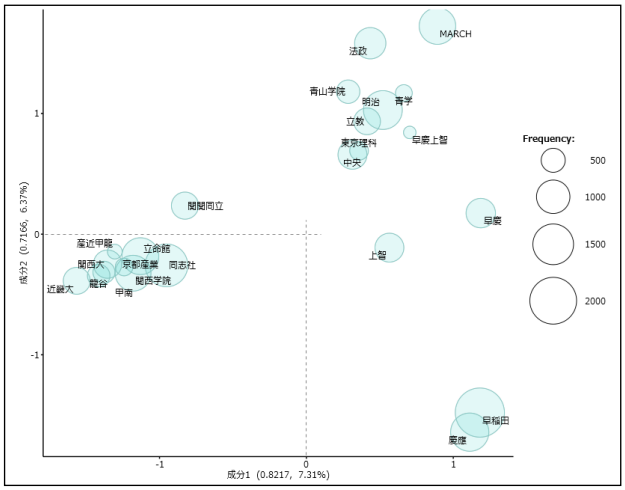

図１　コレスポンデンス分析の結果

3.3. 共起ネットワーク分析

共起ネットワーク分析の結果は図２の通りとなった．こちらでも「早慶のグループ」「MARCH のグループ」「関西の大学のグループ」という３つのグループに大別されている．図２より，「上智大学」が「早慶」と「MARCH」の両方と共起関係をもっている点は興味深い点であった．さらに，関東と関西の大学間の共起関係として「同志社大学と上智大学，立教大学」「立命館大学と立教大学」が見られた．

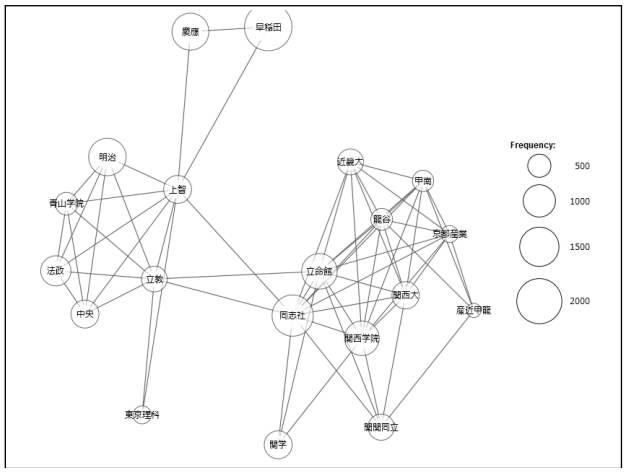

図２　共起ネットワーク分析の結果

4. まとめ

これまでの結果より，本研究の仮説１及び仮説２は概ね支持されている．仮説１では主に「早慶」と「MARCH」のグループによる差異が見られ，特に「MARCH」では相互に一定程度の共起関係があることが明らかとなった．しかしながら，関西の大学では「関関同立」と「産近甲龍」という大学群の名称による明確な差異は確認されなかった．仮説２に対しては，コレスポンデンス分析より関東と関西の大学による差異が見られた．

謝辞

本研究は ROIS-DS-JOINT(課題番号:00032,研究代表者:小舘亮之)の助成を受けた.また,国立情報学研究所の IDR データセット提供サービスにより,ヤフー株式会社から提供の「Yahoo!知恵袋データ(第 3 版)」を利用した.

参考文献

[1]国立情報学研究所：“情報学研究データリポジトリ”(2019/1/10 閲覧),
https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html