日本語版 Wikipedia における量的特徴のカテゴリ間比較
−「良質な記事」に注目して−

大畑　澄子

科学技術の発展が進む現代では一般の人々の間で学術情報・専門知識への要求が高まっている。一般の人々が利用するメディアのなかで Wikipedia は影響力が強く、情報源としての質の高さが求められてきている。そこで既存の記事の改善や新規記事の作成に有用な知見を獲得するため、質の高い記事の特徴を明らかにすることを目標として「良質な記事」と「非・良質な記事」の特徴を計量的に分析しカテゴリ間での比較を行った。

分析対象とする特徴量は長さ、文字数、画像数、外部リンク数、見出し数、編集回数、編集者数、引用注数、参考文献数、閲覧数、表数、数式数である。これらの特徴量は Wikipedia 記事のダンプ・データ、MediaWiki API、ページビュー分析を利用して抽出・集計し算出した。それらの抽出した特徴量を計算し、さらに見出しあたり文字数、異なり編集者数、登録編集者数、参考文献密度の 4 特徴量を求めた。抽出した特徴量は基本集計を行い中央値などの要約統計量を計算し、ウィルコクソンの順位和検定を用いてその差の有意性を検定した。判別分析を行い特徴量による記事の分類可能性を確かめた後、ランダムフォレストを通して分類への影響力の強い特徴量を求めた。

各カテゴリにおける量的特徴の分析においては、全てのカテゴリに共通して記事の長さの方が文字数よりも影響力が強いという結果が得られた。このことから、「良質な記事」の選考においては単純な文字数に加えて文字以外の総合的な情報量が重視されている可能性が示唆された。また、ランダムフォレストを行なった結果、全特徴量のうちもっとも影響力が強い特徴量は引用注数であることが明らかになった。

カテゴリごとにおける量的特徴の分析においては、カテゴリによって影響力が強い特徴量は異なることが明らかになった。天文学、化学の学問分野では参考文献数、科学者・数学者・物理学者、政治的指導者といった人物に関する分野では参考文献密度の影響力がそれぞれ強いという結果になった。こうしたカテゴリ間における特徴量の影響力の違いはカテゴリの持つ分野的特徴が関わっていると考えられる。

今後は、本研究で明らかにした特徴をもとに、「良質な記事」における各特徴量の目標数値を算出する方法を検討することで、質の高い記事作成における具体的な指標を提示することが可能になると考えられる。

（指導教員　芳鐘冬樹）