東京大学大学院
工学系研究科
SCHOOL OF ENGINEERING
THE UNIVERSITY OF TOKYO

新しい再帰的指標によって Wikipedia 編集者と記事の多様性を捉える

１．発表者：

島田　尚
（東京大学大学院工学系研究科システム創成学専攻／東京大学数理・情報教育研究センター（略称：ＭＩセンター）数学基礎教育部門　准教授）

２．発表のポイント：

- ◆ Wikipedia の編集関係ネットワークから編集者の編集傾向と記事の性質とを特徴付ける再帰的指標を提案し、その有効性を示した。
- ◆ 本研究によって、客観的に特徴づけることが難しかった英語版 Wikipedia の編集者と記事の「生態系」を捉えることが可能になった。
- ◆ Wikipedia に代表される巨大なデジタル集合知の場は近年急速に重要性を増しており、今回の成果はそれらの運営と質の向上に資する重要な成果である。今後はより一般の集合知システムや生態系などへの展開が期待される。

３．発表概要：

今年で 20 周年を迎える Wikipedia は参加者による自由な編集を原則として運営されているオンライン百科事典であり、この運営と記事の質を支えている原理については不明な点が多い。

統計物理学を基盤とした国際共同研究グループである東京大学大学院工学系研究科の島田尚准教授、大阪大学の小串典子助教、Central European University の János Kertész 教授、Aalto 大学の Kimmo Kaski 教授らは、Wikipedia の編集関係ネットワークから、編集内容の傾向や記事の性格といった、単なる良し悪しではない性質を客観的に特徴づけることのできる新しい指標を提案した。また、これを用いて英語版 Wikipedia の編集者と記事の「生態系」の構造とダイナミクスを捉えた。

Wikipedia のようなデジタル集合知の場は近年急速に重要性を増しており、今回の成果はそれらの健全な運営と質の向上に資する成果といえる。今後は、Wikipedia 各言語コミュニティ間の差異の検証や、他の集合知システムや生態系への手法の適用を通じて、今回捉えた関係性をさらに明らかにしてゆくことが期待される。

４．発表内容：

現在広く利用されている Wikipedia は、よく知られているように自由な参加者による編集を原則として運営されているオンライン百科事典である。このような自由参加型の集合知システムが安定に運営されまた記事の質が保たれるかについては、悲観的な予想もされてきた。その一方で、創立 20 周年を迎えた現在では Wikipedia は最大の英語コミュニティで 500 万記事を超える規模へと順調に成長しており、また記事の質についても従来型の辞書と同程度という評価もされている。

こうした Wikipedia の成功を支えている重要な仕組みとして、コミュニティ内での議論に基づいて行われている、編集者や記事の評価づけ（ラベルや称号の付与）がある。編集者や記事の数の膨大さからこのような評価を網羅的に行うことは非常に手間のかかるものとなっており、実際に「秀逸な記事」や「良質な記事」などの認定のペースは Wikipedia 全体の成長に追

いついているとは言い難い現状である。このような特徴づけをより客観的・自動的に行うことができればその意義は大きいが、Google の PageRank などに代表される既存の重要性手法では「編集回数が際立っている編集者」や「人気や論争の面で目立つ記事」というような比較的分かりやすい重要度しか評価することができなかった。

このような状況に対して今回、統計物理学を基盤とした国際共同研究グループである東京大学大学院工学系研究科の島田尚准教授、大阪大学の小串典子助教、Central European University の János Kertész 教授、Aalto 大学の Kimmo Kaski 教授らは、編集者の編集内容の傾向（記事の内容についての加筆、記事の体裁や形式の統一、編集合戦への対応、etc.）に対応する「散漫度」と、記事の性質（「良質な記事」、「論争のある記事」、「人気の高い記事」等）に対応する「複雑度」という新しい二つの指標を提案した。これらの指標は、Wikipedia の編集関係のネットワーク（どの編集者がどの記事を編集したかの二部グラフ）を用いて、「複雑度の高い記事のみを編集している編集者は記事コンテンツの加筆の傾向が高い（散漫度が低い）」「記事の内容加筆の傾向が高い編集者による編集を多くうけるほどその記事の複雑性は高い」という自己無撞着な関係を表現する再帰的な計算から定められる（図左）。

今回出版された論文ではまず、こうして求められた「複雑度」を用いる事で、英語版 Wikipedia において「秀逸な記事」や「良質な記事」と認定されている記事を「論争のある記事」や「人気の高い記事」と区別することができることが確かめられた。さらに、既存の重要性指標と組み合わせる事でこれらの性格の異なる記事の違いがうまく捉えられる事を示し、またこれによって記事の特徴の時間発展についてもその多様な振る舞いを明らかにした（図右）。この解析からはさらに、「論争になっている記事であるが加筆には専門的な知識が要求されるもの」など、人の手によるラベリングとの一致だけに留まらない記事の特徴づけが可能になった。

Wikipedia のようなデジタル集合知の場は近年急速に重要性を増しており、今回の成果はそれらの健全な運営と質の向上に資する成果といえる。また、本研究では、これまでの重要性指標では難しかった、単なる良し悪しではない性質を客観的に特徴づけることが可能となり、これにより英語版 Wikipedia の編集者と記事の「生態系」の構造とダイナミクスを捉えることができた。実際に今回の手法は生態系における受粉関係などの相互作用の研究や、経済における国の競争力や生産物の重要性の評価法の研究などと理論的なつながりが深く、これらの問題に通底する「隠れた構造の発見と特徴づけ」という一般的な問題についても貢献するものである。今後は Wikipedia の各言語コミュニティ間の差異の検証を行い、また他の集合知システムや生態系などへ本手法を適用して研究を進める事で、今回明らかになったような集団における多様な役割とその意義についての理解を深めてゆくことが期待できる。

５．発表雑誌：

雑誌名： Scientific Reports vol. 11, 18371 (2021).
論文タイトル：“Ecology of the digital world of Wikipedia”
著者： Fumiko Ogushi, János Kertész, Kimmo Kaski, & Takashi Shimada*
DOI 番号： https://doi.org/10.1038/s41598-021-97755-w
アブストラクト： https://www.nature.com/articles/s41598-021-97755-w

６．問い合わせ先：

東京大学大学院工学系研究科システム創成学専攻／

東京大学数理・情報教育研究センター（略称：ＭＩセンター）数学基礎教育部門　准教授
島田　尚　（しまだ　たかし）

７．添付資料：
（今回の成果の概念図）