信頼される AI システムを支える基盤技術　　2022 年度
2021 年度採択研究代表者　　年次報告書

鹿島 久嗣

京都大学 大学院情報学研究科
教授

人と AI の協働ヒューマンコンピュテーション基盤

主たる共同研究者:
荒井 ひろみ（理化学研究所 革新知能統合研究センター ユニットリーダー）
小山 聡（北海道大学 大学院情報科学研究院 准教授）
森 純一郎（東京大学 大学院情報理工学系研究科 准教授）

研究成果の概要

本研究では、信頼できる人-AI 協働系の設計論の確立を目指し、その信頼性の定義として、従来のコンピュータシステムの信頼性指標を起点としたヒューマンコンピュテーション(H/C)のための指標の定義とこれを達成するための技術開発、また、人-AI 協働系が社会で受容されるための倫理的課題の特定と解決技術の開発、さらに、H/C の活用シナリオとしての、多人数データ解析や創造的課題解決における具体的な技術的課題の解決と実践を目指す。

2 年度目となる今年度は、H/C における信頼性の定義を行うとともに、これを達成するための要素技術の開発を中心に、4 グループで研究を進めた。

まず、H/C 信頼基盤技術グループでは、従来のコンピュータの信頼性の指標の指標として用いられる Reliability（信頼性）、Availability（可用性）、Serviceability（保守性）の観点から、既存の H/C 研究を整理するとともに、H/C における信頼性の論点をまとめ、サーベイ論文を執筆・公開した。また、H/C の実証基盤として用いるコンペティション型多人数データ解析システムの開発を進め、保守性やセキュリティの向上を実現した。

H/C の社会受容グループでは、非専門家に対する AIの判断の説明のあり方について調査を行った。また、人−AI 協働系における公平性について、データ作成におけるバイアスの影響と評価方法、バイアスのあるデータへの対応方法についての検討を行った。特に、多数の人間による意見を統合して結論を出す際に、マイノリティの意見が反映されにくいという問題に対し、これらの影響を適切に補正することによって、公平な意見統合を実現する手法を開発した。

人間参加型機械学習グループでは、少量データから難しい学習課題を達成する要素技術開発を中心に実施した。特に、強化学習問題において、専門家による大量の教示データ作成のボトルネックを解決するために、過去に蓄積されたエピソードデータや、これらをクラウドソーシング等を用いて評価したデータを用いて、効率的な学習を実現する手法を開発した。また、多数の人間が、それぞれのもつ複数の主観的な観点からデータの類似度評価を行ったデータから、これらを解きほぐし、深層学習で利用可能な表現を抽出する手法を開発した。さらに、様々なタスクへの参加を通じて人間がスキルを成長させていくために、スキルの獲得状況を、人間に理解可能な形で推定・提示する方法を開発した。

最後に、H/C による知的創造活動支援グループでは、前年度に収集し、研究の基盤として整備を行なった大規模実データ（学術文献データ、データ解析コンペティションのログデータ、オンラインコミュニケーションデータなど）をもとに、特に科学技術領域における H/C による知的創造活動支援に関する研究を進めた。特に、大規模な学術文献データを人が解釈・理解するための技術として、時系列のトピックモデルや言語生成モデルなどを新たな開発し、成果を発表するとともに公開した。また、論文投稿を行う際に、インパクトに応じて投稿先を推薦する手法を開発した。

【代表的な原著論文情報】

1) Ryosuke Ueda, Koh Takeuchi, Hisashi Kashima. Fair Opinion Aggregation for Voter Attribute Bias. In Proceedings of 6th AAAI/ACM Conference on AI, Ethics, and Society (AIES), 2023.
2) Xiaotian Lu, Jiyi Li, Koh Takeuchi, Hisashi Kashima. Multiview Representation Learning from

Crowdsourced Triplet Comparisons. In Proceedings of the Web Conference (WWW), 2023.
3) Guoxi Zhang, Hisashi Kashima. Behavior Estimation from Multi-Source Data for Offline Reinforcement Learning. In Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI), 2023
4) Guoxi Zhang, Hisashi Kashima. Learning State Importance for Preference-based Reinforcement Learning. Machine Learning, 2022.
5) Ryoma Sato, Makoto Yamada, Hisashi Kashima. Poincare: Recommending Publication Venues via Treatment Effect Estimation. Journal of Informetrics, 2022.
6) Kaname Muto, Satoshi Oyama and Itsuki Noda. Explainable Recommendation Using Knowledge Graphs and Random Walks, In Proceedings of the 6th IEEE Workshop on Human-in-the-Loop Methods and Future of Work in BigData (IEEE HMData), 2022.
7) Masato Ota, Yuko Sakurai and Satoshi Oyama. Coalitional Game Theoretic Federated Learning. In Proceedings of the 21st IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2022.
8) Masato Shinoda, Yuko Sakurai and Satoshi Oyama. Sample Complexity of Learning Multi-value Opinions in Social Networks. In Proceedings of the 24th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA), 2022.
9) Masanao Ochi, Masanori Shiro, Junichiro Mori, Ichiro Sakata, Predictive analysis of multiple future scientific impacts by embedding a heterogeneous network, PLOS ONE, 17(9), 2022.
10) Takahiro Miura, Kimitaka Asatani and Ichiro Sakata, Revisiting the uniformity and inconsistency of slow-cited papers in science, Journal of Informetrics, 17(1), 2022.
11) Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori and Ichiro Sakata, Dynamic Structured Neural Topic Model with Self-Attention Mechanism, In Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL), 2023.
12) Masaru Isonuma, Junichiro Mori and Ichiro Sakata, Differentiable Instruction Optimization for Cross-Task Generalization, In Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL), 2023.
13) Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori and Ichiro Sakata, SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation, In Findings of the Association for Computational Linguistics: ACL 2023 (Findings of ACL), 2023.