様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ５ 月 ３１ 日現在

機関番号：１２１０２
研究種目：基盤研究(B)（一般）
研究期間：2019～2023
課題番号：１９Ｈ０４２１５

研究課題名（和文）A scalable privacy-preserving information retrieval system based on federated optimization, on-device intelligence and semantic matching

研究課題名（英文）A scalable privacy-preserving information retrieval system based on federated optimization, on-device intelligence and semantic matching

研究代表者

于　海涛（Yu, Haitao）

筑波大学・図書館情報メディア系・准教授

研究者番号：３０７５１０５２

交付決定額（研究期間全体）：（直接経費）　　12,200,000円

研究成果の概要（和文）：本プロジェクトは、プライバシーを考慮した高度な情報検索技術の開発を目指している。このプロジェクトの終了までに、以下の主要な視点を通じて顕著な成果を上げた：フェデレーテッドラーニングと差分プライバシーを組み合わせた新しい情報アクセス手法の提案、直接的評価指標の最適化、大規模言語モデルに基づく検索結果の再ランキング、個人化を考慮した会話型情報検索手法の提案。その結果、20篇以上の学会論文と10篇以上の雑誌論文を発表した。さらに、ランキング学習のための使いやすいオープンソースプロジェクトの研究開発を行い、代表的なランキング学習の手法を複数実装している。

研究成果の学術的意義や社会的意義
Our research achievements would deepen the understanding of privacy-preserving information seeking that goes beyond information retrieval (IR). By releasing the source codes and collections, we encourage the entire IR community to improve the research of privacy-preserving IR towards new stages.

研究成果の概要（英文）：This project aims to initiate research into privacy-preserving information retrieval (IR). Throughout the lifetime of this project, we made remarkable achievements via the following main aspects: novel ways of combining federated learning and differential privacy for privacy-preserving information access, direct optimization of evaluation metrics, effective integration of LLMs for result re-ranking, incorporating personalized context for conversational information seeking. As a result, we published more than 20 conference papers and 10 journal papers. Moreover, we are also maintaining an open-source project for IR named as PTRanking, which includes many representative ranking methods based on neural networks. Overall, it is reasonable to say that the successful accomplishment of this project will bring new insights into the development of privacy-preserving IR techniques.

研究分野：情報検索

キーワード：Federated Learning　Large Language Model　Conversational IR　Generative IR　Personalization　On-device Learning　User Modeling　Neural Tree Ensemble

様　式　C－１９、F－１９－１（共通）

１．研究開始当初の背景

Information Retrieval (IR) systems has been playing an important role in our daily information access. Significant research efforts have been made on different aspects of IR. However, accurately and efficiently providing satisfactory results to users while preserving privacy is far from being resolved. The traditional IR models are trained and improved to gain insights into the overall user population. Unfortunately, the data to derive such insights is personal and sensitive, which might give rise to catastrophic consequences, even if the system collecting such data has resolved to 'do no evil'. Recently, differential privacy has increasingly gained popularity as a benchmark of protecting the privacy of data subjects while providing aggregated query responses over databases. Federated learning provides an alternative method for performing distributed privacy-protecting machine learning. To the best of our knowledge, how to achieve effective privacy-preserving IR has not been well studied yet, and no practical privacy-preserving IR system was ever deployed, neither proposed. Specifically, **the key fundamental challenges include**: (1) *Most successes, if not all, of the current ranking models rely on enormous amounts of labeled data which is available beforehand*. Moreover, they learn each task in isolation and from scratch based on incremental model updates using stochastic gradient descent. In result, these models are inherently data-hungry and time-consuming, which hinders their application in privacy-preserving IR. (2) *In response to an input query, the current IR platforms rely entirely on the high-performance clusters running on the cloud to afford the computational tasks*, such as the representation of context information and the analysis of a user's intent. This setting significantly complicates the protection of user privacy, since lots of private and sensitive data are kept server-side. (3) *The current IR systems follow the same fashion of 'one size fits all', namely training and applying a single ranking model across all users, which is insufficient to satisfy different users' preference*. Thus providing user-specific search results without privacy breach while interacting with a huge number of users is another challenge that has to be solved.

２．研究の目的

The critical demand of sufficient early preparations and new solutions towards privacy-protecting IR motivate us to address privacy-preserving IR in a novel way. By proposing this project, we aim to explore how to achieve effective privacy-preserving IR by focusing on on-device user understanding, federated optimization and semantic matching. The key outcomes of this project will be a suite of effective machine learning methods for privacy-preserving IR, such as on-device intent detection and semantic matching approches based on deep neural networks, which will be published as either conference papers or journal papers. Meanwhile, The research results (e.g., source codes and test collections) will be disseminated via a website, facilitating the advancement of privacy-preserving IR research to new stages.

３．研究の方法

The proposed methods for addressing privacy-preserving IR are designed and developed from the following three aspects.

Aspect-1: On-device intent detection. The key idea works like this: given the query 'Disneyland ticket', instead of performing an encryption-decryption process like the transaction and payment security solutions, the on-device intent-detection model is designed to identify the semantic meaning by incorporating the context information (e.g., the search history and the user's profile) and represent this user's information need as a real-valued vector. Based on the semantic matching framework, search results will be returned. Moreover, the intent-detection model that resides on an individual's device, collects the search behaviors and the data about the user over time, and performs a self-guided learning to optimize the ability of intent identification.

Aspect-2: Federated optimization framework. When limited to a single user, the local data might sometimes be too sparse to be useful in the self-guided on-device learning. For instance, some users are likely to repeatedly search for a single type of information. To overcome this problem, the core idea is to aggregate locally computed model updates (e.g., gradient vectors) across the entire user population in a federated fashion.

Aspect-3: Semantic matching framework. Based on the training data consisting of search requests and documents, we jointly learn two models that respectively transform raw search requests and documents into vector representations. The recent embedding techniques and large language models (LLMs) will be used, and all the vector representations will belong to

the same latent high-dimensional space.

4．研究成果

The research achievements can be summarized from the following three groups, which echoes the aforementioned three aspects.

Group-1: User understanding. Throughout the lifetime of this project, we believe that conversational information seeking (CIS) would play an important role in the future, which extends the classic search to a conversational nature. Our focused framework (Figure 1) for CIS accounts for users' personas [1]. It assumes that there is a personal text knowledge base (PTKB), which consists narrative sentences providing personal information about the users. Specifically, the developed system consists of the following key modules. (1) Statement ranking: given the context of the conversation and the current user utterance, this module returns a ranked list of PTKB statements based on their relevance, which reflects the user's persona; (2) Passage ranking: given the context of the conversation, the current user utterance, and the PTKB statements, this module is responsible for retrieving a ranked list of passages from the document collection; (3) Response generation: this module returns the answer text as a response to the user. In particular, the response should be a generative or abstractive summary of the relevant passages. The experimental results based on the dataset released by TREC iKAT track show that: (1) For PTKB statement ranking, our method achieves the best performance in terms of MRR on the set of iKAT organizers' assessments. It also shows superior performance over the baseline based on GPT-4. This indicates that a fusion of multiple LLMs is a promising choice when tackling problems of this kind. (2) For passage ranking, on one hand, one of our proposed strategies is able to achieve comparable performance as Llama2-based baseline. On the other hand, our analysis indicates that the way of incorporating PTKB statements for personalized retrieval matters, where a direct concatenation is not recommended. (3) For response generation, our proposed method is able to generate grounded and natural personalized responses, and is comparable to the top-tier LLM-based baseline.

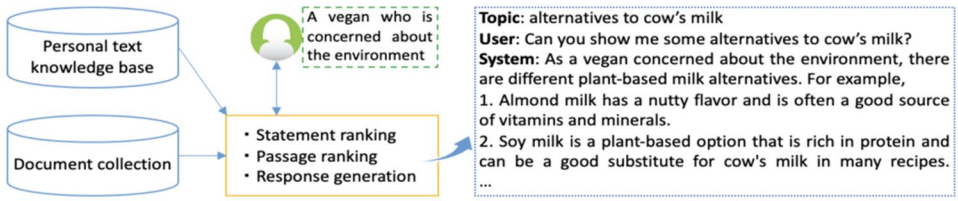

Figure 1. An illustration of the focused CIS system

Group-2: Federated optimization. Besides a direct application of the traditional federated learning algorithm (namely a decentralized learning manner where multiple clients collaboratively train a global model while keeping their data local), we tried to explore how to improve it. For example, we explored a new federated learning framework [9] in the shuffle model and a simple protocol extended from existing work.

Group-3: Semantic matching. To cope with this challenge, we explored from different directions and achieved promising outcomes, such as incorporating personalized context [1], diversified result ranking [10], taking into account the temporal features [8, 13], direct metric optimization [6, 10], and deploying LLMs for advanced ranking [1, 2, 4]. Take the proposed direct metric optimization for diversified document ranking for example, we proposed a novel framework through direct metric optimization for search result diversification based on the score-and-sort strategy. The experimental results on benchmark collections show that the proposed method achieves significantly improved performance over the state-of-the-art results.

5. Highlighted Publications[1]:

[1] Hai-Tao Yu, Lingzhen Zheng, Kaiyu Yang, Sumio Fujita and Hideo Joho. Towards Incorporating Personalized Context for Conversational Information Seeking. Proceedings of the International Workshop on Information Retrieval's Role in RAG Systems (IR-RAG), 2024.

[1] Please refer to the following link for the entire list of publications:

https://github.com/ii-research/PPIR/blob/master/Accomplishments/ResearchAccomplishments.md

[2] Kaiyu Yang, Lingzhen Zheng, Hai-Tao Yu, Sumio Fujita and Hideo Joho. University of Tsukuba Team at the TREC 2023 Deep Learning Track. Proceedings of The Thirty-Second Text REtrieval Conference (TREC 2023), 2023.

[3] Haonan Tan, Kaiyu Yang and Hai-Tao Yu. An In-depth Comparison of Neural and Probabilistic Tree Models For Learning-to-Rank. Proceedings of The 46th European Conference on Information Retrieval (ECIR), 468–476, 2024.

[4] Lingzhen Zheng, Kaiyu Yang, Hai-Tao Yu, Sumio Fujita and Hideo Joho. University of Tsukuba Team at the TREC 2023 Interactive Knowledge Assistance Track. Proceedings of the Thirty-Second Text REtrieval Conference (TREC 2023), 2023.

[5] Yun Gao, Hai-Tao Yu, Xin Kang and Fuji Re. TUA1 at the TREC 2019: Deep Learning Track. Proceedings of the 28th Text Retrieval Conference (TREC), 2019.

[6] Muramoto Naoki and Hai-Tao Yu. Deep Metric Learning Based on Rank-sensitive Optimization of Top-k Precision. Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM), 2161-2164, 2020.

[7] Jiexin Wang, Adam Jatowt and Masatoshi Yoshikawa. Event Occurrence Date Estimation based on Multivariate Time Series Analysis over Temporal Document Collections. The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 398-407, 2021.

[8] Lirong Zhang, Hideo Joho and Hai-Tao Yu. Semantic Modelling of Document Focus-time for Temporal Information Retrieval. In Companion Proceedings of the Web Conference 2022 (WWW '22 Companion): 12th International Workshop on Temporal Web Analytics (TempWeb), 2022.

[9] Ruixuan, Liu, Yang Cao, Hong, Chen, Ruoyang, Guo and Masatoshi, Yoshikawa. FLAME: Differentially Private Federated Learning in the Shuffle Model. The 35th AAAI Conference on Artificial Intelligence (AAAI-21), 8688-8696, 2021.

[10] Hai-Tao Yu. Optimize What You EvaluateWith: Search Result Diversification Based on Metric Optimization. The 36th AAAI Conference on Artificial Intelligence, 36(9), 10399-10407, 2022

[11] Hai-Tao Yu, Degen Huang, Fuji Ren and Lishuang Li. Diagnostic Evaluation of Policy-Gradient-Based Ranking. Electronics, 2022.

[12] Hai-Tao Yu, Piryani Rajesh, Jatowt Adam, Inagaki Ryo, Hideo Joho and Kim, Kyoung-Sook. An In-Depth Study on Adversarial Learning-to-Rank. Information Retrieval Journal, 26(1), 2023.

[13] Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa and Yi Cai. BiTimeBERT: Extending Pre-Trained Language Representations with Bi-Temporal Information. The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 812-821, 2023.

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　3件／うちオープンアクセス　1件）

１．著者名 Yu Hai-Tao、Piryani Rajesh、Jatowt Adam、Inagaki Ryo、Joho Hideo、Kim Kyoung-Sook	４．巻 26
２．論文標題 An in-depth study on adversarial learning-to-rank	５．発行年 2023年
３．雑誌名 Information Retrieval Journal	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10791-023-09419-0	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 MURAMATSU Naoya、YU Hai-Tao、SATOH Tetsuji	４．巻 E106.D
２．論文標題 Combining Spiking Neural Networks with Artificial Neural Networks for Enhanced Image Classification	５．発行年 2023年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 252～261
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2021EDP7237	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kang Xin、Shi Xuefeng、Wu Yunong、Ren Fuji	４．巻 14
２．論文標題 Active Learning With Complementary Sampling for Instructing Class-Biased Multi-Label Text Emotion Classification	５．発行年 2023年
３．雑誌名 IEEE Transactions on Affective Computing	６．最初と最後の頁 523～536
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TAFFC.2020.3038401	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Hai-Tao Yu; Degen Huang; Fuji Ren; Lishuang Li	４．巻 11
２．論文標題 Diagnostic Evaluation of Policy-Gradient-Based Ranking	５．発行年 2022年
３．雑誌名 Electronics	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/electronics11010037	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Kang Xin、Shi Xuefeng、Wu Yunong、Ren Fuji	４．巻 1
２．論文標題 Active learning with complementary sampling for instructing class-biased multi-label text emotion classification	５．発行年 2020年
３．雑誌名 IEEE Transactions on Affective Computing	６．最初と最後の頁 1〜1
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TAFFC.2020.3038401	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計21件（うち招待講演　0件／うち国際学会　13件）

１．発表者名 Lirong Zhang、Hideo Joho、Hai-Tao Yu
２．発表標題 Semantic Modelling of Document Focus-Time for Temporal Information Retrieval
３．学会等名 The 12th International Workshop on Temporal Web Analytics (TempWeb)（国際学会）
４．発表年 2022年

１．発表者名 Xin Kang、Rongyu Dou、Haitao Yu
２．発表標題 TUA1 at eRisk 2022: Exploring Affective Memories for Early Detection of Depression
３．学会等名 The Thirteenth Conference and Labs of the Evaluation Forum (CLEF 2022)（国際学会）
４．発表年 2022年

１．発表者名 Lirong Zhang、Hideo Joho、Hai-Tao Yu
２．発表標題 Selectively Expanding Queries and Documents for News Background Linking
３．学会等名 The 31st ACM International Conference on Information and Knowledge Management (CIKM2022)（国際学会）
４．発表年 2022年

１．発表者名
Shuyuan Zheng, Yang Cao, Masatoshi Yoshikawa, Huizhong Li, Qiang Yan

２．発表標題
FL-Market: Trading Private Models in Federated Learning

３．学会等名
2022 IEEE International Conference on Big Data (Big Data)（国際学会）

４．発表年
2022年

１．発表者名
Hai-Tao Yu

２．発表標題
Optimize What You EvaluateWith: Search Result Diversification Based on Metric Optimization

３．学会等名
The 36th AAAI Conference on Artificial Intelligence（国際学会）

４．発表年
2022年

１．発表者名
Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa

２．発表標題
Event Occurrence Date Estimation based on Multivariate Time Series Analysis over Temporal Document Collections

３．学会等名
The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)（国際学会）

４．発表年
2021年

１．発表者名
Fumiyuki Kato, Yang Cao, Masatoshi Yoshikawa

２．発表標題
Preventing Manipulation Attack in Local Differential Privacy using Verifiable Randomization Mechanism

３．学会等名
The IFIP Annual Conference on Data and Applications Security and Privacy（国際学会）

４．発表年
2021年

１．発表者名
Yusuke Kubono, Xin Kang, Fuji Ren, Shun Nishide

２．発表標題
Prediction and Generation of Multiple Complex Drawing Figures From Partial Drawing Sequences

３．学会等名
The IEEE 13th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)（国際学会）

４．発表年
2021年

１．発表者名
F Ding, Xin Kang, Yunong Wu, Fuji Ren

２．発表標題
TUA1 at the NTCIR-16 DialEval-2 Task

３．学会等名
The 16th NTCIR Conference on Evaluation of Information Access Technologies

４．発表年
2022年

１．発表者名
Naoya Muramatsu, Hai-Tao Yu

２．発表標題
Combining Spiking Neural Network and Artificial Neural Network for Enhanced Image Classification

３．学会等名
第13回データ工学と情報マネジメントに関するフォーラム

４．発表年
2021年

１．発表者名
Muramoto Naoki, Hai-Tao Yu

２．発表標題
Deep Metric Learning Based on Rank-sensitive Optimization of Top-k Precision

３．学会等名
Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)（国際学会）

４．発表年
2020年

１．発表者名
Hai-Tao Yu

２．発表標題
Neural Learning-to-Rank

３．学会等名
言語処理学会第27回年次大会ワークショップ：クリエイティブAIが会話型AIと出会うとき

４．発表年
2021年

１．発表者名
Ruixuan, Liu and Yang Cao and Hong, Chen and Ruoyang, Guo and Masatoshi, Yoshikawa

２．発表標題
FLAME: Differentially Private Federated Learning in the Shuffle Model

３．学会等名
The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)（国際学会）

４．発表年
2021年

１．発表者名
Ruixuan, Liu and Yang Cao and Masatoshi, Yoshikawa and Hong, Chen

２．発表標題
FedSel: Federated SGD Under Local Differential Privacy with Top-k Dimension Selection

３．学会等名
Database Systems for Advanced Applications (DASFAA 2020)（国際学会）

４．発表年
2020年

１．発表者名
門田 尚之，康 シン，西出 俊，任 福継

２．発表標題
ロボット教師における特定分野のQAシステムの構築

３．学会等名
2020年電子情報通信学会総合大会

４．発表年
2020年

１．発表者名
寺尾 渉吾，康 シン，西出 俊，任 福継

２．発表標題
ロボット教師における特定人物の音声の自動生成

３．学会等名
2020年電子情報通信学会総合大会

４．発表年
2020年

１．発表者名
房登 淳平，康 シン，西出 俊，任 福継

２．発表標題
単語の感情属性を活かしたロボット教師の表情生成

３．学会等名
2020年電子情報通信学会総合大会

４．発表年
2020年

１．発表者名
山崎 直登，康 シン，西出 俊，任 福継

２．発表標題
ロボット教師による特定分野の記述式問題の判定について

３．学会等名
2020年電子情報通信学会総合大会

４．発表年
2020年

１．発表者名
Yun Gao, Hai-Tao Yu, Xin Kang, Fuji Ren

２．発表標題
TUA1 at the TREC 2019: Deep Learning Track

３．学会等名
Proceedings of the 28th Text Retrieval Conference（国際学会）

４．発表年
2019年

１．発表者名
Yasunobu Sumikawa, Adam Jatowt, Antoine Doucet, Jean-Philippe Moreux

２．発表標題
Large Scale Analysis of Semantic and Temporal Aspects in Cultural Heritage Collection's Search

３．学会等名
Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2019)（国際学会）

４．発表年
2019年

１．発表者名
Hai-Tao Yu, Lingzhen Zheng, Kaiyu Yang, Sumio Fujita, Hideo Joho

２．発表標題
Towards Incorporating Personalized Context for Conversational Information Seeking

３．学会等名
The International Workshop on Information Retrieval's Role in RAG Systems (IR-RAG)

４．発表年
2024年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

PPIR
https://github.com/ii-research/PPIR

Learning to Rank in PyTorch
https://ptl2r.github.io

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	吉川　正俊 (Yoshikawa Masatoshi) (30182736)	大阪成蹊大学・データサイエンス学部・教授 (34437)	

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	康　シン (Kang Xin) (80777350)	徳島大学・大学院社会産業理工学研究部（理工学域）・助教 (16101)	
研究分担者	Ａｄａｍ　Ｊａｔｏｗｔ (Jatowt Adam) (00415861)	京都大学・情報学研究科・特定准教授 (14301)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
オーストリア	University of Innsbruck			
中国	Dalian University of Technology			