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　４ 年　 ５ 月 ３１ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１１９７９

研究課題名（和文）Efficient Query Processing for Learning-based Data Management

研究課題名（英文）Efficient Query Processing for Learning-based Data Management

研究代表者
肖　川（Xiao, Chuan）

大阪大学・情報科学研究科・准教授

研究者番号：１０６４３９００

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では、機械学習に基づくデータマネジメントについてのクエリ処理を対象として、効率的なクエリ処理手法の開発に関する研究を行った。特に、埋め込みベクトルに対するクエリを効率的に処理するため、二つの解決策を開発した。1つ目は、2値の高次元ベクトルに対して、ハミング距離制約のある類似クエリを効率的に回答する。2つ目は、実数値の高次元ベクトルに対して、階層的なグラフ構造を用いて、近似最近傍探索を行う。また、機械学習の述語を含むクエリの処理を研究し、カーディナリティ推定により高速なクエリプランを生成する手法を開発した。システムのプロトタイピングと評価を行い、ソフトウェアのソースコードを公開した。

研究成果の学術的意義や社会的意義
本研究の成果は、機械学習に基づくデータマネジメントの実践的な手法を提供し、次世代データマネジメントシステムの開発に貢献する。最先端のデータベース技術を進展させ、機械学習、自然言語処理、コンピュータビジョンなどの関連研究分野やマーケティング、医療などの応用での技術開発に強い推進力を与える。また、日本のコンピュータサイエンスにおける威信を高め、海外の研究グループとのコラボレーションを促進することにも貢献する。

研究成果の概要（英文）：We addressed several fundamental problems of query processing for learning-based data management. We developed two solutions to efficient processing of queries on embedding vectors: the first works for binary high-dimensional vectors and efficiently returns answers for similarity search and join queries with Hamming distance constraints; the second handles approximate nearest neighbor search for real-valued high-dimensional vectors by utilizing hierarchical graph structures. We studied the processing of queries with learning-based predicates and developed methods that generate fast query plans through cardinality estimation. We performed system prototyping and evaluation, and released the source codes of our software at GitHub. The outcome of this project provides practical methods for learning-based data management and contributes to the development of next-generation data management systems.

研究分野：情報科学

キーワード：クエリ処理　ML for DB　高次元データ　類似検索

１．研究開始当初の背景

Recent decades have witnessed a rapid proliferation of data. In the meanwhile various machine learning (ML) models have been developed for a wide range of challenging tasks. A recent trend in the database (DB) research area is to utilize ML techniques for DB problems, such as entity matching, near-duplicate detection, and query optimization. Despite some early attempts from the DB research community, two key questions regarding ML for DB are yet to be addressed. First, the execution speeds are not efficient for large datasets, hence limiting the applicability of ML methods on large datasets. Second, existing ML for DB methods were mainly developed for specific applications (e.g., healthcare and product specification) and thus it is difficult to transfer them to other application scenarios, even if the underlying models are exactly the same. Consequently, it is necessary to develop novel data management methods to cope with the efficiency issue from a generic perspective to benefit a wide range of DB problems and their applications.

２．研究の目的

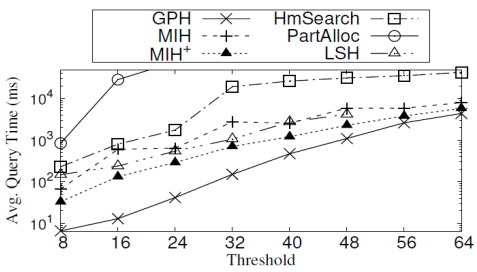
This project aims at addressing fundamental problems and designing novel methods to improve the efficiency of data management systems that integrate ML methods. Three tasks are performed in this project: (1) developing efficient query processing for embedding vectors, (2) developing generic blocking techniques to process queries with ML-based predicates, and (3) prototyping a system that integrates the proposed methods and evaluating it on large-scale datasets. This project is characterized as an exploration in the next-generation core DB technology which features ML techniques in a variety of tasks.

３．研究の方法

We develop efficient indexing and query processing techniques for embedding vectors, utilizing the pigeonhole principle for filtering and hierarchical graph structures for indexing. To develop generic blocking techniques, we focus on the processing of conjunctive queries of similarity predicates generated through active learning, and design solutions that generate fast query plans using accurate cardinality estimation.

４．研究成果

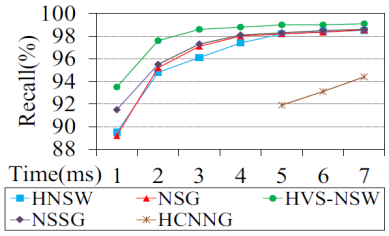
A)　(Task 1) We studied the problem of query processing for binary high-dimensional vectors. We developed a solution that efficiently returns answers for similarity search and join queries with Hamming distance constraints [1]. Answering such queries efficiently plays an important role in web search, image retrieval, and scientific databases. We generalized the pigeonhole principle and designed our query processing algorithm. As shown in Figure 1, the proposed method (GPH) delivers very promising query processing performance and it is 4 - 10 times faster than existing solutions.



(d) GIST, Query Processing Time

Figure 1. Evaluation of Task 1 (binary vectors).

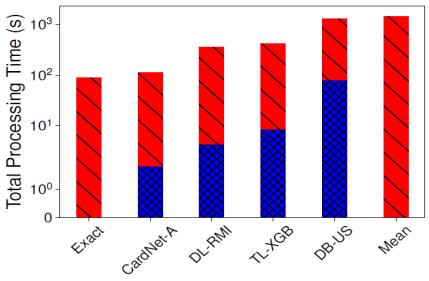
B) (Task 1) By utilizing hierarchical graph structures, we proposed an indexing approach to approximate nearest neighbor search for real-valued high-dimensional vectors [2]. Such query is used in various applications of ML, such as image & video retrieval and recommender systems. Figure 2 shows that under the same query processing time constraint, our method (HVS) improves recall rates by 3% - 10% when compared to existing ones. In addition, it spends less indexing time than existing methods.



(b) Gist, $K = 1$

Figure 2. Evaluation of Task 1 (real-valued vectors).

C) (Task 2) We targeted blocking rules of conjunctive similarity predicates generated through active learning. We modeled the processing of these rules as a query optimization problem and developed a deep learning-based method that generates fast query plans through cardinality estimation [3]. As shown in Figure 3, the proposed method (CardNet) is up to one order of magnitude faster than the traditional method (DB-US) that employs sampling techniques for cardinality estimation. We also developed a method that specializes in real-valued high-dimensional vectors and improves the accuracy of cardinality estimation [4].



(c) Time, IMDB-Movie

Figure 3. Evaluation of Task 2.

D) (Task 3) We finished system prototyping and evaluated the methods developed in this project using large-scale datasets. We released the source codes of our software [5, 6, 7, 8] at GitHub. We reported our discoveries in this project and gave tutorials [9, 10] at top-tier academic venues.

[1] J. Qin, C. Xiao, Y. Wang, W. Wang, X. Lin, Y. Ishikawa, and G. Wang. Generalizing the Pigeonhole Principle for Similarity Search in Hamming Space. TKDE, 33(2): 489-505, 2021.

[2] K. Lu, M. Kudo, C. Xiao, and Y. Ishikawa. HVS: Hierarchical Graph Structure Based on Voronoi Diagrams for Solving Approximate Nearest Neighbor Search. PVLDB, 15(2): 246-258, 2021.

[3] Y. Wang, C. Xiao, J. Qin, X. Cao, Y. Sun, W. Wang, and M. Onizuka. Monotonic Cardinality Estimation of Similarity Selection: A Deep Learning Approach. SIGMOD, 1197-1212, 2020.

[4] Y. Wang, C. Xiao, J. Qin, R. Mao, M. Onizuka, W. Wang, R. Zhang, and Y. Ishikawa. Consistent and Flexible Selectivity Estimation for High-Dimensional Data. SIGMOD, 2319-2327, 2021.

[5] https://github.com/chuanxiao1983/GPH

[6] https://github.com/chuanxiao1983/HVS

[7] https://github.com/chuanxiao1983/CardNet

[8] https://github.com/chuanxiao1983/SelNet

[9] J. Qin, W. Wang, C. Xiao, and Y. Zhang. Similarity Query Processing for High-Dimensional Data. PVLDB, 13(12): 3437-3440, 2020.

[10] J. Qin, W. Wang, C. Xiao, Y. Zhang, and Y. Wang. High-Dimensional Similarity Query Processing for Data Science. KDD, 4062-4063, 2021.

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　9件／うち国際共著　5件／うちオープンアクセス　9件）

１．著者名 Kejing Lu, Mineichi Kudo, Chuan Xiao, Yoshiharu Ishikawa	４．巻 15(2)
２．論文標題 HVS: Hierarchical Graph Structure Based on Voronoi Diagrams for Solving Approximate Nearest Neighbor Search	５．発行年 2021年
３．雑誌名 Proceedings of the VLDB Endowment	６．最初と最後の頁 246-258
掲載論文のDOI（デジタルオブジェクト識別子） 10.14778/3489496.3489506	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 －

１．著者名 Yuyang Dong, Chuan Xiao, Hanxiong Chen, Jeffrey Xu Yu, Kunihiro Takeoka, Masafumi Oyamada, Hiroyuki Kitagawa	４．巻 30(2)
２．論文標題 Continuous Top-k Spatial-Keyword Search on Dynamic Objects	５．発行年 2021年
３．雑誌名 The VLDB Journal	６．最初と最後の頁 141-161
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00778-020-00627-4	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 Jianbin Qin, Chuan Xiao, Yaoshu Wang, Wei Wang, Xuemin Lin, Yoshiharu Ishikawa, Guoren Wang	４．巻 33(2)
２．論文標題 Generalizing the Pigeonhole Principle for Similarity Search in Hamming Space	５．発行年 2021年
３．雑誌名 IEEE Transactions on Knowledge and Data Engineering	６．最初と最後の頁 489-505
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TKDE.2019.2899597	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 Jianbin Qin, Wei Wang, Chuan Xiao, Ying Zhang	４．巻 13
２．論文標題 Similarity Query Processing for High-Dimensional Data	５．発行年 2020年
３．雑誌名 Proceedings of the VLDB Endowment	６．最初と最後の頁 3437-3440
掲載論文のDOI（デジタルオブジェクト識別子） 10.14778/3415478.3415564	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 Satoshi Koide, Chuan Xiao, Yoshiharu Ishikawa	4．巻 13
2．論文標題 Fast Subtrajectory Similarity Search in Road Networks under Weighted Edit Distance Constraints	5．発行年 2020年
3．雑誌名 Proceedings of the VLDB Endowment	6．最初と最後の頁 2188-2201
掲載論文のDOI（デジタルオブジェクト識別子） 10.14778/3407790.3407818	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 －

1．著者名 Jianbin Qin, Chuan Xiao, Sheng Hu, Jie Zhang, Wei Wang, Yoshiharu Ishikawa, Koji Tsuda, Kunihiko Sadakane	4．巻 29(4)
2．論文標題 Efficient Query Autocompletion with Edit Distance-based Error Tolerance	5．発行年 2020年
3．雑誌名 The VLDB Journal	6．最初と最後の頁 919-943
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00778-019-00595-4	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 小出 智士, 肖 川, 石川 佳治	4．巻 J103-D
2．論文標題 道路ネットワーク上の軌跡データに対する圧縮索引	5．発行年 2020年
3．雑誌名 電子情報通信学会論文誌 D	6．最初と最後の頁 393-402
掲載論文のDOI（デジタルオブジェクト識別子） 10.14923/transinfj.2019DET0001	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 －

1．著者名 Sheng Hu, Chuan Xiao, Yoshiharu Ishikawa	4．巻 27
2．論文標題 Scope-aware Code Completion with Discriminative Modeling	5．発行年 2019年
3．雑誌名 IPSJ Journal of Information Processing	6．最初と最後の頁 469-478
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjjip.27.469	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 －

1．著者名 Jing Zhao, Yoshiharu Ishikawa, Lei Chen, Chuan Xiao, Kento Sugiura	4．巻 E102-D
2．論文標題 Building Hierarchical Spatial Histograms for Exploratory Analysis in Array DBMS	5．発行年 2019年
3．雑誌名 IEICE Transactions on Information and Systems	6．最初と最後の頁 788-799
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2018DAP0020	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

〔学会発表〕　計25件（うち招待講演　0件／うち国際学会　11件）

1．発表者名 Koji Matsuda, Yuya Sasaki, Chuan Xiao, Makoto Onizuka
2．発表標題 FedMe: Federated Learning via Model Exchange
3．学会等名 SIAM International Conference on Data Mining (SDM)（国際学会）
4．発表年 2022年

1．発表者名 Misato Horiuchi, Yuya Sasaki, Chuan Xiao, Makoto Onizuka
2．発表標題 JupySim: Jupyter Notebook Similarity Search System
3．学会等名 International Conference on Extending Database Technology (EDBT)（国際学会）
4．発表年 2022年

1．発表者名 Sheng Hu, Ichigaku Takigawa, Chuan Xiao
2．発表標題 Edit-Aware Generative Molecular Graph Autocompletion for Scaffold Input
3．学会等名 Deep Learning on Graphs: Method and Applications (DLG-AAAI)（国際学会）
4．発表年 2022年

１．発表者名
胡晟，瀧川一学，肖川

２．発表標題
深層生成モデルを用いた編集を意識した分子グラフ補完

３．学会等名
データ工学と情報マネジメントに関するフォーラム (DEIM)

４．発表年
2022年

１．発表者名
松本和人，肖川，鬼塚真

２．発表標題
学習型索引を用いた時系列データ検索の高速化

３．学会等名
データ工学と情報マネジメントに関するフォーラム (DEIM)

４．発表年
2022年

１．発表者名
河越淳，董于洋，野澤拓磨，肖川

２．発表標題
Attention GANを用いたテーブルデータの欠測値補完

３．学会等名
データ工学と情報マネジメントに関するフォーラム (DEIM)

４．発表年
2022年

１．発表者名
川本孝太朗，伊藤竜一，佐々木勇和，肖川，鬼塚真

２．発表標題
結合カーディナリティ推定の中間結果を利用した結合順最適化

３．学会等名
データ工学と情報マネジメントに関するフォーラム (DEIM)

４．発表年
2022年

１．発表者名
三宅康太，佐々木勇和，肖川，鬼塚真

２．発表標題
統合型データベースにおける適応的2相ロックに基づく分散トランザクション制御

３．学会等名
データ工学と情報マネジメントに関するフォーラム（DEIM）

４．発表年
2022年

１．発表者名
松田光司，佐々木勇和，肖川，鬼塚真

２．発表標題
モデル構造の自動チューニングを用いたパーソナライズド連合学習手法

３．学会等名
データ工学と情報マネジメントに関するフォーラム（DEIM）

４．発表年
2022年

１．発表者名
池田悠人，三宅康太，肖川，鬼塚真

２．発表標題
機械学習によるトランザクション処理性能の網羅的な評価

３．学会等名
データ工学と情報マネジメントに関するフォーラム（DEIM）

４．発表年
2022年

１．発表者名
Jianbin Qin, Wei Wang, Chuan Xiao, Ying Zhang, Yaoshu Wang

２．発表標題
High-Dimensional Similarity Query Processing for Data Science

３．学会等名
ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)（国際学会）

４．発表年
2021年

１．発表者名
Yaoshu Wang, Chuan Xiao, Jianbin Qin, Rui Mao, Makoto Onizuka, Wei Wang, Rui Zhang, Yoshiharu Ishikawa

２．発表標題
Consistent and Flexible Selectivity Estimation for High-Dimensional Data

３．学会等名
ACM SIGMOD International Conference on Management of Data (SIGMOD)（国際学会）

４．発表年
2021年

１．発表者名
Yuyang Dong, Kunihiro Takeoka, Chuan Xiao, Masafumi Oyamada

２．発表標題
Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach

３．学会等名
IEEE International Conference on Data Engineering (ICDE)（国際学会）

４．発表年
2021年

１．発表者名
Simin Yu, Kuntian Zhang, Chuan Xiao, Xianyu Bao, Joshua Zhexue Huang, Mark Junjie Li

２．発表標題
BTGAN: Training GAN with Balanced Triplet Loss and Two-Branch Architecture

３．学会等名
International Joint Conference on Neural Networks (IJCNN)（国際学会）

４．発表年
2021年

１．発表者名
伊藤竜一，佐々木勇和，肖川，鬼塚真

２．発表標題
Non-Autoregressiveモデルによる高速で安定したカーディナリティ推定

３．学会等名
データ工学と情報マネジメントに関するフォーラム (DEIM)

４．発表年
2021年

１．発表者名
松田光司，堀敬三，佐々木勇和，肖川，鬼塚真

２．発表標題
FedMe：モデル交換に基づく連合学習手法

３．学会等名
データ工学と情報マネジメントに関するフォーラム（DEIM）

４．発表年
2021年

１．発表者名
堀内美聡，山崎翔平，佐々木勇和，肖川，鬼塚真

２．発表標題
計算ノートブック類似検索のための高速な検索アルゴリズム

３．学会等名
データ工学と情報マネジメントに関するフォーラム（DEIM）

４．発表年
2021年

１．発表者名
胡晟，瀧川一学，肖川

２．発表標題
深層生成モデルを用いた分子グラフ自動補完

３．学会等名
データ工学と情報マネジメントに関するフォーラム（DEIM）

４．発表年
2021年

１．発表者名
Yaoshu Wang, Chuan Xiao, Jianbin Qin, Xin Cao, Yifang Sun, Wei Wang, Makoto Onizuka

２．発表標題
Monotonic Cardinality Estimation of Similarity Selection: A Deep Learning Approach

３．学会等名
ACM SIGMOD International Conference on Management of Data (SIGMOD)（国際学会）

４．発表年
2020年

１．発表者名
三宅 康太，涌田 悠佑，佐々木 勇和，肖 川，鬼塚 真

２．発表標題
P2P型データ統合アーキテクチャにおけるチケットベース手法を用いた分散トランザクション制御

３．学会等名
データ工学と情報マネジメントに関するフォーラム (DEIM)

４．発表年
2020年

１．発表者名
高 明敏，肖 川，石川 佳治

２．発表標題
トライ木及びGMMに基づく略語のフルネームのスケーラブルな推測手法

３．学会等名
データ工学と情報マネジメントに関するフォーラム (DEIM)

４．発表年
2020年

１．発表者名
胡 晟，馬 強，肖 川

２．発表標題
多様化軌跡を効率検索するための統合クエリパラダイム

３．学会等名
データ工学と情報マネジメントに関するフォーラム (DEIM)

４．発表年
2020年

１．発表者名
Makoto Onizuka, Yusuke Wakuta, Yuya Sasaki, Chuan Xiao

２．発表標題
Distributed Transaction Management for P2P-based Update Propagation

３．学会等名
Workshop on Software Foundations for Data Interoperability (SFDI)（国際学会）

４．発表年
2019年

１．発表者名
Sheng Hu, Chuan Xiao, Jianbin Qin, Yoshiharu Ishikawa, Qiang Ma

２．発表標題
Autocompletion for Prefix-Abbreviated Input

３．学会等名
ACM SIGMOD International Conference on Management of Data (SIGMOD)（国際学会）

４．発表年
2019年

１．発表者名
Daichi Amagata, Takahiro Hara, Chuan Xiao

２．発表標題
Dynamic Set kNN Self-Join

３．学会等名
IEEE International Conference on Data Engineering (ICDE)（国際学会）

４．発表年
2019年

〔図書〕　計1件

１．著者名	４．発行年
Lu Qin, Wenjie Zhang, Ying Zhang, You Peng, Hiroyuki Kato, Wei Wang, Chuan Xiao	2020年
２．出版社	５．総ページ数
Springer	193
３．書名	
Software Foundations for Data Interoperability and Large Scale Graph Data Analytics - 4th International Workshop, SFDI 2020, and 2nd International Workshop, LSGDA 2020, held in Conjunction with VLDB 2020, Tokyo, Japan, September 4, 2020, Proceedings	

〔産業財産権〕

〔その他〕

研究発表の備考：
大阪大学 ビッグデータ工学講座 鬼塚研究室
http://www-bigdata.ist.osaka-u.ac.jp/ja/paper/
名古屋大学 情報学研究科 データベース研究室（石川研究室）
https://www.db.is.i.nagoya-u.ac.jp/ja/research/publications/
Chuan Xiaoのホームページ
https://sites.google.com/site/chuanxiao1983/publication
Chuan XiaoのDBLPページ
https://dblp.org/pid/57/4384-1.html

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

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			
オーストラリア	ニューサウスウェールズ大学	メルボルン大学	シドニー工科大学	
中国	香港科技大学	深セン大学	深セン計算科学研究院	他1機関