データ駆動科学におけるデータ照合技法に関する研究

Study on data matching technology for data driven science

川島英之
慶應義塾大学環境情報学部

１．研究目的

計算機アーキテクチャの発展とセンサ技術の向上に伴いデータ処理を高速化する需要は日々高まっている。この中で整列技法は重要な要素技術である。近年並列性を利用することにより整列を高性能化する多数の技法が提案されてきた。本研究の目的はメニーコアアーキテクチャを利用することにより高性能化される整列技法を提案することだった。従来型の整列技法に加えて、深層学習などの関数にもとづき整列をおこなう、いわゆる learned sort も対象にした。

２．研究成果の内容

並列 sort について、PARADIS[1], Regions sort[2]を比較実験しており、PARADIS のスケラビリティに着目した。PARADIS はほかの並列 sort 法と比較しても高い並列性を示すことが論文に示されているが、性能劣化を引き起こす挙動に関する解析は不十分である。本研究では PARADIS を再実装し、特定の場合において PARADIS の処理が逐次的に行われてしまうケースを明らかにし、そのケースに対応するための新しい PARADIS の手法，適応的分割法を提案する。適応的分割法は、従来の PARADIS では不可能であった Repair フェーズ内のバケット中の操作を並列化することを可能にした。適応的分割法を導入した PARADIS は実験の結果、最大で 57% の性能向上を実現した。

[1] Cho, M., Brand, D., Bordawekar, R., Finkler, U., Kulandaisamy, V., & Puri, R. (2015). PARADIS: an efficient parallel algorithm for in-place radix sort. Proceedings of the VLDB Endowment, 8(12), 1518–1529.
[2] Obeya, O., Kahssay, E., Fan, E., & Shun, J. (2019). Theoretically-efficient and practical parallel in-place radix sorting. Annual ACM Symposium on Parallelism in Algorithms and Architectures, 213–224.

３．学際共同利用が果たした役割と意義

学際共同利用において大規模並列分散環境について理解をすることができ、新しいアルゴリズムの構想を具体化することができた。この意味において学際共同利用は本研究に有益だった。

４．今後の展望

今後は機械学習をベースとした learned sort の高性能化に取り組むことを考えている。整列法は重要な課題であるため、その高性能化には引き続き重要であると考えている。特に範囲検索を行うような場合には通常 index を作成するが、index には膨大なメモリ空間が必要になってしまう。これを learned algorithm を用いることで、問題を解決したいと考えている。

５．成果発表

(1) 学術論文

(2) 学会発表

(3) その他

使用計算機	使用計算機に○	配分リソース※	
		当初配分	追加配分
Cygnus			
Oakforest-PACS	○	24576	
※配分リソースについてはノード時間積をご記入ください。			