グリッドコンピューティング環境における
データベース処理に関する研究

（課題番号 3106）

平成１５年度～平成１７年度特定研究費
（一般研究費）
研究成果報告書

平成１９年１０月

研究代表者　　北 上　 始

（広島市立大学大学院情報科学研究科・教授）

はしがき

本研究は，平成１５年度から平成１７年度までの３年間にわたって交付された特定研究費（一般研究費）による研究課題「グリッドコンピューティング環境におけるデータベース処理に関する研究」（課題番号 3106）の成果を取りまとめたものである．

近年，企業・公共機関・教育機関・家庭などの至る所でインターネットに接続された数多くのパーソナルコンピュータが活用されてきたことが引き金となり，最近，インターネット上に接続されている複数台のパソコンを「超スーパーコンピュータ」として利用する技術が注目されている．この技術は，単一コンピュータの性能限界を安価に解消する現実的な技術であり，グリッドコンピューティングと呼ばれている．グリッドコンピューティング環境では，企業等で代表される各組織内に設置されている複数台のパーソナルコンピュータ（以下，ＰＣと呼ぶ）をクラスタと呼び，各組識間を結ぶコンピュータネットワークを用いて各クラスタ同士が結合されている．本研究では，グリッドコンピューティング環境を実験室内で仮想的に構築し，仮想的なグリッドコンピューティング環境下でデータベースからの知識発見を支援するシステムをどのように構成するのかについて研究を行った．具体的には，（１）配列データマイニング処理の高機能化の方法，（２）複数クラスタ上でその処理を実現する方法などについて研究し，それらの方法の有用性について検討を行った．

本研究は，研究分担者，その他の関係者との共同研究であり，ご協力頂いた関係各位に心からお礼を申し上げると共に，本研究成果として得られた知見が今後の情報社会になんらかの形で役立つことを切望するものである．

平成１９年１０月

研究代表者　北上　始

（広島市立大学大学院情報科学研究科・教授）

目 次

研究組織 -1-

研究発表 -2-

研究成果概要 -6-

研究成果 -8-

(1) 北上　始：日本のデータベース研究最前線，ＤＮＡ配列の謎を解き明かす配列データマイニング，月刊 DB マガジン，翔泳社，pp.172-173，2006 年 5 月． -9-

(2) 加藤 智之，北上 始，森 康真，田村 慶一，黒木 進：極小かつ非冗長な可変長ワイルドカード領域を持つ頻出パターンの抽出，電子情報通信学会論文誌 D「データ工学特集号」，Vol.J90, No.2, pp.281-291，2007 年 2 月． -11-

(3) 加藤 智之，北上 始，森 康真，田村 慶一，黒木 進： 極小な可変長ワイルドカード領域をもつ頻出配列パターンの抽出，日本データベース学会論文誌 (DBSJ Letters), Vol.5, No.1, pp.117-120, 2006 年 6 月． -23-

(4) Makoto Takaki, Keiichi Tamura, and Hajime Kitakami: Dynamic Load Balancing Technique for Modified PrefixSpan on a Grid Environment with Distributed Worker Model, Proceedings of The 2006 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'06 & RTCOMP'06), Vol. Ⅱ, Las Vegas, Nevada, USA, pp.895-901, June 26-30, 2006. -27-

(5) Tomoyuki Kato, Hajime Kitakami, Makoto Takaki, Keiichi Tamura, Yasuma Mori, and Susumu Kuroki:: Extraction for Frequent Sequential Patterns with Minimum Variable-Wildcard Regions, Proceedings of The 2006 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'06 & RTCOMP'06), Vol. Ⅱ, Las Vegas, Nevada, USA, pp.825-831, June 26-30, 2006.. . -36-

(6) 田村慶一、岩木稔、高木允、北上始：PC クラスタにおける混合整数計画問題の並列処理とその性能評価，情報処理学会論文誌：数理モデル化と応用，Vol. 46, No.SIG 17(TOM 13), pp.56-69，2005 年 12 月． -45-

(7) Makoto TAKAKI, Keiichi TAMURA, Toshihide SUTOU and Hajime KITAKAMI: New Dynamic Load Balancing for Parallel Modified PrefixSpan with Distributed Worker Paradigm, Proc. of International Special Workshop on Databases for Next Generation Researchers in Memoriam of Prof. Kambayashi (SWOD2005), pp.96-99, Tokyo in Japan, April 2005, also to be appeared in Proc. of ICDE Workshops, IEEE Computer Society, p.1243, 2005. -60-

(8) 高木允，田村慶一，周藤俊秀，北上始：並列 Modified PrefixSpan 法の並列化と動的負荷分散手法，情報処理学会論文誌：数理モデル化と応用，Vol. 46, No.SIG 10 (TOM 12), pp.138-152，2005 年 6 月． -65-

(9) Makoto TAKAKI, Keiichi TAMURA, Toshihide SUTOU and Hajime KITAKAMI: Dynamic Load Balancing for Parallel Modified PrefixSpan, Proceedings of The 2004 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'04), Vol.1, Las Vegas, Nevada, USA, pp.352-358, June 21-24, 2004. -81-

(10) 塔野 薫隆，北上 始，田村 慶一，森 康真，黒木 進：Modified PrefixSpan 法を用いた頻出正規パターンの抽出をめざして，日本データベース学会論文誌(DBSJ Letters)，Vol.3，No.1，pp.61-64，2004 年 6 月． -90-

(11) 周藤 俊秀，田村 慶一，森 康真，北上 始：並列 Modified PrefixSpan 法の設計と実装，日本データベース学会論文誌(DBSJ Letters), Vol.2. No.3, pp.25-28, 2003 年 12 月． -94-

(12) Toshihide SUTOU, Keiichi TAMURA, Yasuma MORI, Hajime KITAKAMI: Design and Implementation of Parallel Modified PrefixSpan Method, Proc. of the fourth International Symposium on High Performance Computing, Lecture Notes in Computer Science (LNCS),Springer-Verlag, Vol.2327, pp.412-422, October 2003. -98-

研 究 組 織

研究代表者： 北 上　始（広島市立大学大学院情報科学研究科・教授）

研究分担者： 黒 木　進（広島市立大学大学院情報科学研究科・准教授）
研究分担者： 森　康 真（広島市立大学大学院情報科学研究科・助教）
研究分担者： 田村 慶一（広島市立大学大学院情報科学研究科・助教）

研 究 経 費（配 分 額）

（金額単位：千円）

	直 接 経 費	間 接 経 費	合 計
平成１５年度	1,250	0	1,250
平成１６年度	800	0	800
平成１７年度	900	0	900
総　計	2,950	0	2,950

研 究 発 表

ア. 解説記事

1．北上　始：日本のデータベース研究最前線，ＤＮＡ配列の謎を解き明かす配列データマイニング，月刊 DB マガジン，翔泳社，pp.172-173，2006 年 5 月.

イ. 学会誌等

1. Toshihide SUTOU, Keiichi TAMURA, Yasuma MORI, Hajime KITAKAMI: Design and Implementation of Parallel Modified PrefixSpan Method, Proc. of the fourth International Symposium on High Performance Computing, Lecture Notes in Computer Science (LNCS),Springer-Verlag, Vol.2327, pp.412-422, October 2003.
2. 周藤 俊秀，田村 慶一，森 康真，北上 始：並列 Modified PrefixSpan 法の設計と実装，日本データベース学会論文誌(DBSJ Letters), Vol.2. No.3, pp.25-28, 2003 年 12 月.
3. 塔野 薫隆，北上 始，田村 慶一，森 康真，黒木 進：Modified PrefixSpan 法を用いた頻出正規パターンの抽出をめざして，日本データベース学会論文誌 (DBSJ Letters)，Vol.3，No.1，pp.61-64，2004 年 6 月.
4. Makoto TAKAKI, Keiichi TAMURA, Toshihide SUTOU and Hajime KITAKAMI: Dynamic Load Balancing for Parallel Modified PrefixSpan, Proceedings of The 2004 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'04), Vol.1, Las Vegas, Nevada, USA, pp.352-358, June 21-24, 2004.
5. 高木允，田村慶一，周藤俊秀，北上始：並列 Modified PrefixSpan 法の並列化と動的負荷分散手法，情報処理学会論文誌：数理モデル化と応用，Vol. 46, No.SIG 10 (TOM 12), pp.138-152，2005 年 6 月.
6. Makoto TAKAKI, Keiichi TAMURA, Toshihide SUTOU and Hajime KITAKAMI: New Dynamic Load Balancing for Parallel Modified PrefixSpan with Distributed Worker Paradigm, Proc. of International Special Workshop on Databases for Next Generation Researchers in Memoriam of Prof. Kambayashi (SWOD2005), pp.96-99, Tokyo in Japan, April 2005, also to be appeared in Proc. of ICDE Workshops, IEEE Computer Society, p.1243, 2005.

7. 田村慶一、岩木稔、高木允、北上始：PC クラスタにおける混合整数計画問題の並列処理とその性能評価，情報処理学会論文誌：数理モデル化と応用，Vol. 46, No.SIG 17(TOM 13), pp.56-69，2005 年 12 月.

8. Tomoyuki Kato, Hajime Kitakami, Makoto Takaki, Keiichi Tamura, Yasuma Mori, and Susumu Kuroki: Extraction for Frequent Sequential Patterns with Minimum Variable-Wildcard Regions, Proceedings of The 2006 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'06 & RTCOMP'06), Vol. Ⅱ, Las Vegas, Nevada, USA, pp.825-831, June 26-30, 2006.

9. Makoto Takaki, Keiichi Tamura, and Hajime Kitakami: Dynamic Load Balancing Technique for Modified PrefixSpan on a Grid Environment with Distributed Worker Model, Proceedings of The 2006 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'06 & RTCOMP'06), Vol. Ⅱ, Las Vegas, Nevada, USA, pp.895-901, June 26-30, 2006.

10. 加藤 智之，北上 始，森 康真，田村 慶一，黒木 進： 極小な可変長ワイルドカード領域をもつ頻出配列パターンの抽出，日本データベース学会論文誌 (DBSJ Letters), Vol.5, No.1, pp.117-120, 2006 年 6 月

11. 加藤 智之，北上 始，森 康真，田村 慶一，黒木 進：極小かつ非冗長な可変長ワイルドカード領域を持つ頻出パターンの抽出，電子情報通信学会論文誌，データ工学特集号，Vol.J90, No.2, pp.281-291，2007 年 2 月.

ウ. 口頭発表(査読付き)

1. 石橋 和枝, 高木 允, 周藤 俊秀, 田村 慶一, 北上 始: 並列 Modified PrefixSpan 法におけるチェックポイント/リスタートの実装と性能評価, 第 6 回 IEEE 広島学生シンポジウム, pp.298-301, 2004 年 12 月.
2. 井上 俊治, 北上 始, 黒木 進, 森 康真, 田村 慶一 : 異種データベース統合のためのスキーマ間の対応付けに関する研究, 第 6 回 IEEE 広島学生シンポジウム, pp.309-312, 2004 年 12 月.
3. 坂本 尚子, 森 康真, 北上 始 : 木構造で管理されている異種アンケートの横断的分析, 第 6 回 IEEE 広島学生シンポジウム, pp.290-293, 2004 年 12 月.
4. 札埜 有香, 田村 慶一, 森 康真, 北上 始 : 遺伝的プログラミングによる調停グラフの交差数減少とその並列分散処理, 電子情報通信学会・第 16 回データ工学ワークショップ(DEWS2005), 電子情報通信学会データ工学研究専門委員会, Online Proceedings (http://www.ieice.org/iss/de/DEWS/), 2005 年 3 月.
5. 田村慶一, 高木允, 北上始, 札埜有香 : 遺伝的プログラミングによる調停グラフの交差数減少の並列分散処理とそのグリッド環境下における性能評価, 先進的計算基盤システムシンポジウム (SACSIS2005), 情報処理学会, pp.258-259, 2005 年 5 月.
6. 劉 強, 北上 始, 黒木 進, 田村 慶一, 森 康真 : 画像データベースにおける索引構造の効率化, 電子情報通信学会・第 16 回データ工学ワークショップ (DEWS2005), 電子情報通信学会データ工学研究専門委員会, Online Proceedings (http://www.ieice.org/iss/de/DEWS/), 2005 年 3 月.
7. 周藤 俊秀, 高木 允, 田村 慶一, 北上 始: グリッド環境下における Modified PrefixSpan 法の並列処理とその動的負荷分散方式, 先進的計算基盤システムシンポジウム(SACSIS2005), 情報処理学会, pp.161-168, 2005 年 5 月.
8. Shigetaka TONO, Hajime KITAKAMI, Keiichi TAMURA, Yasuma MORI, Susumu KUROKI: Efficiently Mining Sequence Patterns With Variable-Length Wildcard Regions Using An Extended Modified PrefixSpan Method, The 13th Annual International Conference on Intelligent Systems for Molecular Biology(ISMB2005), Poster No.G22, p.147, Detroit in USA, June 2005.
9. 高木允, 田村慶一, 北上始 : 並列 Modified PrefixSpan 法のグリッド化とその動的負荷分散手法, FCS/MPS シンポジウム, 名古屋, pp.241-248, 2005 年.
10. 田村慶一, 高木允, 北上始 : 遺伝的プログラミングによる調停グラフ交差数減少のための並列分散処理, FCS/MPS シンポジウム, 名古屋, pp.271-278, 2005 年.

11. Hajime KITAKAMI, Makoto TAKAKI, Keiichi TAMURA, Tomoyuki KATO, and Susumu KUROKI: The Minimum Generalization of Flexible Sequence Patterns Extracted by Prefix-Projected Pattern-Growth Approach, the 3rd RECOMB Comparative Genomics Satellite Workshop, Poster, Dublin in Ireland, September 2005.
12. 高木允，田村慶一，北上始：グリッド環境下での分散型ワーカモデルを用いた Modified PrefixSpan 法の動的負荷分散方式，電子情報通信学会・第 17 回データ工学ワークショップ(DEWS2006)，電子情報通信学会データ工学研究専門委員会，Online Proceedings（http://www.ieice.org/iss/de/DEWS/），2005 年 3 月.
13. 加藤智之，北上始，森康真，田村慶一，黒木進：極小な可変長ワイルドカード領域を持つ頻出配列パターンの抽出，電子情報通信学会・第 17 回データ工学ワークショップ (DEWS2006)，電子情報通信学会データ工学研究専門委員会，Online Proceedings（http://www.ieice.org/iss/de/DEWS/），2005 年 3 月.
14. 加藤 智之，森 康真，荒木 康太郎，黒木 進，北上 始：可変長配列パターン抽出法におけるギブスサンプリングを用いた不要パターンの除去方式，電子情報通信学会・第 18 回データ工学ワークショップ(DEWS2007)，電子情報通信学会データ工学研究専門委員会，Online Proceedings（http://www.ieice.org/iss/de/DEWS/），2007 年 3 月.
15. 荒木 康太郎，田村 慶一，加藤 智之，黒木 進，北上 始：曖昧検索に基づく最小汎化パターンの抽出法，電子情報通信学会・第 18 回データ工学ワークショップ(DEWS2007)，電子情報通信学会データ工学研究専門委員会，Online Proceedings，2007 年 3 月.

研 究 成 果 概 要

本研究では，グリッドコンピューティング環境を実験室内で仮想的に構築し，仮想的なグリッドコンピューティング環境下でデータベースからの知識発見を支援するシステムをどのように構成するのかについて明らかにするために，(A) 配列データマイニング処理の高機能化の方法，(B) 複数クラスタ上で配列データマイニング処理を実現する方法，などについて研究を行った．そして，以下の研究成果が得られた．

■配列データマイニング処理の高機能化の方法

（1）生命情報科学分野におけるアミノ酸配列や遺伝子配列などの分子配列データベースに着目し，可変長ワイルドカード領域やあいまい文字などの表現を含む正規表現の頻出パターンを抽出する方法の研究を行った．この正規表現の頻出パターン抽出問題は，顧客の購買履歴データベースを対象にした従来の時系列データマイニングでは研究されていなかったため，この研究は，データマイニングを高機能化する研究として位置付けられる．この高機能化の研究の中で，可変長ワイルドカード領域を抽出する方法を実装することにより，従来の手法では見落されていた頻出パターンを抽出することができるようになった．

■複数クラスタ上で配列データマイニング処理を実現する方法

この研究で扱う配列データマイニング処理で生成される列挙木（マイニング木とも呼ばれる）は平衡木ではなく，どの節点についてもそれを頂点とする部分木の深さを予め見積もることができないという問題がある．この問題点を踏まえて，以下の動的な負荷分散方法を開発した．

（1）マスタワーカモデルを応用し，１クラスタが６４台の PC（パーソナルコンピュータ）環境で，良好な台数効果を得る動的負荷分散方法を開発することができた．我々はこの方法をマスタ・タスク・ステイル法と読んでいる．この傾向は，６４台規模まで PC を増加させても，大きな性能低下が見られなかった．また，このマスタ・タスク・ステイル法の汎用性を確認するために，混合整数計画法の並列化に適用した．この結果，良好な台数効果が得られることがわかった．特に，超線形加速（例えば１６台の PC で１台の PC に比べて３０倍の性能向上）の効果を実測できたことは興味深い．

（2）マスタワーカモデルに比べてスケーラビリティのある分散型ワーカモデルを応用し，１クラスタが１００台の PC 環境で，良好な台数効果を得る

動的負荷分散方法を開発することができた．我々はこの動的負荷分散方法をキャッシュベースドランダム・ステイル法と呼んでいる．

（3） 3クラスタをダミーネットで結合した仮想的なグリッドコンピューティング環境を構築し，この環境下で分散型ワーカモデルを応用し，ダミーネットの遅延に左右されにくい動的負荷分散方法を開発した．我々はこれをキャッシュベースドマルチキャスト・ステイル法と呼んでいる．なお，クラスタ内の動的負荷分散はマスタ・タスク・ステイル法を用いており，ダミーネットは，クラスタ間に遅延機能を備えたプログラムを搭載した PC で実現している．性能評価の結果，キャッシュベースドマルチキャスト・ステイル法は，従来提案されている，ランダムステイル法，マルチキャスト法，キャッシュベースドランダム・ステイル法に比べて，ダミーネットの遅延に左右されにくいことがわかった．

今後の課題については以下のとおりである．

（1） 配列データマイニング処理の高機能化として，あいまい文字表現を含む配列パターンを抽出する方法の研究が残されている．また，ギブスサンプリングとの関係についても多くの検討が残されている．

（2） 配列データマイニングにより抽出された配列パターンを視覚表示し，利用者の知識発見を支援する方法の検討が残されている．

（3） グリッドコンピューティング環境で見られるヘテロな大規模計算機環境において，現在までに提案されている動的負荷分散方法の限界について見極めるための検討が残されている．

（4）グリッドコンピューティング環境ではクラスタや PC などの計算機資源が計算途中に動的に変化することはなかった．資源が動的に変化する環境や PC の性能が不均一な環境などでの動的な負荷分散の方法に関する検討が残されている．また，移動体のみで一時的な無線ネットワークを形成するアドホックネットワーク環境における動的負荷分散方法の検討も今後重要になると思われる．

（5） 我々が扱った配列データマイニング処理を関係データベース操作言語ＳＱＬやオブジェクト指向データベース言語などの実用的なデータベース操作言語にどのように統合するのかという研究が残されている．