平成 22 年度 RICC 利用報告書

課題名（タイトル）：

理研サイネスデータベースを用いた大規模分散処理

利用者氏名：　豊田　哲郎
所属　　：　横浜研究所　生命情報基盤研究部門

１．　本課題の研究の背景、目的、関係するプロジェクトとの関係

理研サイネスは，ライフサイエンス分野の様々なデータベースを標準化されたデータ形式で格納し，データベースを統合的に編纂，公開する為のフレームワークである．生命情報基盤研究部門では，この理研サイネスを運用するにあたり，単に各データベースを公開するだけでなく，データ統合により得られる様々な付加価値を見出す研究開発を進めている．研究成果として得られる理研内外に向けたデータ公開サービスには，複数のデータベースにまたがるデータのつながりをグラフィカルに見せるウエブページの提供や，データベース横断的な検索サービスの提供があり．これを支える内部処理として，データレコード間につながりをデータベース横断的に調べるクローリングと呼ぶプロセスがある．現在，上記公開サービスの提供のために定常的なクローリングを必要としているデータレコードが1000 万以上存在する．これらデータレコードに対するクローリングには膨大な計算リソースが必要である．本プロジェクトは，このクローリング処理にRICC の計算リソースを用いることで，処理に必要な総所用時間の短縮を目的とするものである．

２．　具体的な利用内容、計算方法

理研サイネスのクローリング・ジョブの実行時間は，そのジョブに含まれるデータレコード数と比例する関係がある．今回は，これらデータレコードの集合を単純に分割し，大規模分散化による高速化手法を採用した（図1）．

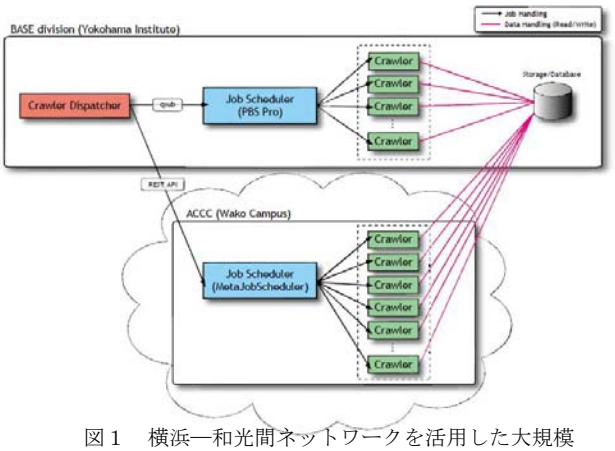

図１　横浜―和光間ネットワークを活用した大規模分散クローラ実行システムのアーキテクチャ

分散化されたジョブは，横浜側に設置されたストレージにデータの読み込み及び結果の書き出しを行うことで計算処理が進行する．

図 1 に示すように，現在利用可能な計算リソースとして，生命情報基盤研究部門が既に持っているリソースと，RICC から割り当てられるリソースがある．それぞれのリソース・ロケーションには，それぞれローカル・スケジューラが備わっている．昨年度までに，これらに対してジョブ割り当てするためのディスパッチャを開発し，さらに横浜研究所に配置されたデータベースを和光の計算リソースから利用するためのネットワーク環境の整備は完了している．

今年度はこれらの成果を元に，日々規模が増大する理研サイネスに格納された全データを対象とした定常的なクローリングの実行を進め，理研サイネスの安定的なデータ公開を図った．

３．　結果

クローリング対象であるデータレコード数は，昨年度末で約 450 万件であった．今年度末は 1000 万件以上にわたり、倍増している．また，１つのデータレコードのクローリングの速度に大きく関与するパラメータとして，当該データレコードと他のデータレコード間の関係を記述するセマンティックリンクと呼ばれるリ

ンクの数がある．このセマンティックリンクの数も，昨年度は 1 データレコードあたり 2.0 リンクから 9.3 リンクへと増大した．更に，理研サイネスの発展的拡張のために，クローリングプログラムで生成するデータの複雑さが増大し，１データレコードあたりのデータ処理量が約 3.5 倍に増大した．

一方、今年度はクローリングプログラムの改良を行うことで、処理に要する時間を短縮に成功した．具体的には，データ量や利用した計算資源の規模が異なるための昨年度の結果と単純比較はできないが，昨年度は毎秒 24 データレコード程度であったスループットを，今年度は横浜研究所にあるリソースを併用して毎秒 12 データレコード程度と，半分程度の低下に抑えることができた．今年度末現在，全データレコードのクローリングに要する時間は 10 日程度である．

４． まとめ

今回，RICC と理研サイネスを広域イーサネットで接続し，RICCを用いた大規模分散化によるクローリング処理の高速化手法を採用し，検証を行った．結果，理研サイネスの運用上問題の無い速度で，定常的にクローリングを実行させることができた．

５． 今後の計画・展望

今後理研サイネスでは，さらなるデータ数の増加が見込まれる．また、理研サイネスのサービスの拡張に伴い，クローリングプログラムで処理すべき内容も追加されることが予想される．このような状況下でも，なるべく遅延無くデータを提供できるよう，技術検討を行う予定である．

６． RICC の継続利用を希望の場合は，これまで利用した状況（どの程度研究が進んだか，研究においてどこまで計算出来て，何が出来ていないか）や，継続して利用する際に行う具体的な内容

理研サイネスは，理研データベースの公開を目的とし，継続運用されるべきシステムである．本課題は、理研サイネスの内部処理であるクローリングを対象に，システムの実現運用研究を実証的に推進するものである．これまでのRICCを用いた検証により，クローリングについては安定運用実現に一定の見通しがつき，大規模分散化への方法論を得た．クローリングの高速化への対応については，各分散計算機に割り当てるジョブに含まれるデータレコード数を調整することで実現させることを検討している．検討に必要なデータは，クローラを実行させつつ蓄積しているところであり，今後はこれを基に実環境に沿って最適化を行う．また，同時実行数を増やした場合は同時にストレージへのアクセスが生じるため，ストレージが本来持つ性能が発揮できない問題も生じている．この問題にもより根本的な技術改良を検討することで，大きな同時実行数での実行が可能となるよう研究開発を進める．

７． 利用研究成果が無かった場合の理由

本課題にて RICC を利用しているクローリング・ジョブは理研サイネスを構成する定常的なサービスであり、研究的な成果を得ることを目的としたプログラムではないため．