平成 23 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

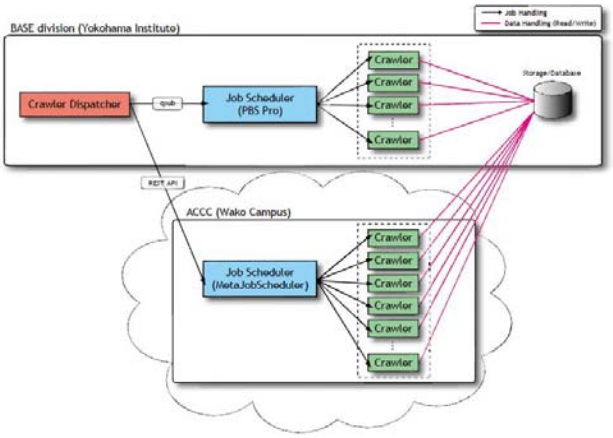
利用者氏名：豊田 哲郎
理研での所属研究室名：横浜研究所 生命情報基盤研究部門

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，データクラス毎に各種フォーマット（TSV・RDF・GFF 等）でつながり情報を表現したファイルの提供（BioLOD，http://biolod.org/）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）．



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，定常的にクローリングを実行できる体制となっている．

今年度はこれらの成果を元にデータ間での優先順位を付けたクローリングの仕組み作りを進め，理研サイネスの安定的なデータ公開を図った．

３． 結果

今年度末の時点で，クローリング対象であるデータレコード数は約 1060 万件，データレコード間またはデ

ータレコードと定数とのつながりを表すデータは約 7080 万件である．また，データレコードは約 28 万のデータクラスによって分類されている．理研サイネスが公開されてから 3 年目となり，データレコードの増加は公開当初と比較して緩和されたが，今年度はデータクラス毎に提供するファイル種別が増加し，またきめ細かいアクセス制御を行うこととなったため，クローリング時の計算量は格段に増加している．また，今年度は新規データの登録に代わって既存データが更新されることが多くなった．新規データ登録は通常非公開状態で行われ，クローリング実行後に公開されることが多いが，データ更新は公開状態のままで行われる．このため，最新のクローリング結果を提供するために処理にはより即時性が求められることになった．

そこで，今年度はクローリングプログラムの改良を行うとともに，新規・更新データを重点的にクローリングする仕組みを作成した．具体的には，①新規・更新データを対象とした即時クローリング，②更新頻度の高いデータを対象とした毎晩のクローリング，③全データを対象とした定期クローリング，という 3 種を併用して運用することにした．この結果，昨年度よりも短いタイムラグでユーザーに最新のクローリング結果ファイルを提供できるようになった．なお今年度末現在，全データレコードのクローリングに要する時間は 15 日程度である．

４． まとめ

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，クローリング結果を提供できるようになった．

５． 今後の計画・展望

今年度の研究で，新規・更新データを優先的に処理する仕組みを実現した．しかし，理研サイネスのクローリングでは複数段階にわたるデータ間のリンク情報を探索しており，リンクの末端にあるデータに関しては定期クローリングを待たなければデータが更新されない状況にある．今後は，処理自体の速度向上と併せ，上記の問題の解決に向けて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，クローリングの安定運用については一定の見通しが付いた．今後は，今年度取り組んだ新規・更新データを対象とした優先的クローリングを改良し，最小限のタイムラグでのデータ提供を目指し検討を行う予定である．

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．