平成 24 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国際標準規格（セマンティックウエブ形式）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 1200 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．

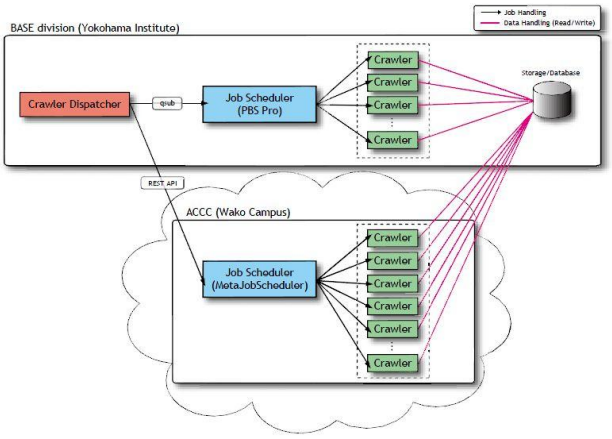

図 1 REST API を用いたジョブ・ステータスのモニタリングおよび投入

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．一昨年度までに，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.また，昨年度は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．

今年度は，さらに増加するデータを処理するため，これまでの実績を元にデータ探索範囲の見直しや公開時の表示上の工夫を行うことで，クローリング・ジョブの所用時間短縮を図った．

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 1230 万件（昨年度 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 7960 万件（昨年度 7080 万件）と増加している．昨年度の時点で全データレコードのク

ローリングに要する時間は約 15 日間であったが，データ量の増加に伴い，従来の方法ではこの速度を維持することが困難になった．

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，ウエブ画面上での表示方法の変更やデータセット毎のクロール深度の設定により，理研サイネスを閲覧するユーザーの利益を損なうことなくクローリング時間を短縮することを目指した．

まずは，表示方法の変更について説明する．従来の理研サイネスでは，各データに対して予めクローリングプログラムの作成した静的なコンテンツを表示していた．この際，クローリングプログラムではセマンティックウエブで結ばれた最大 3 段階先までのデータを探索していた．今年度はこの方法を変更し，最大 2 段階先までのデータを探索したコンテンツをクローリングプログラムで作成しておき，それより先の段階のデータを閲覧したいユーザーに対しては，表示時に Ajax 等の技術を用いて動的にコンテンツを取得し提供する仕組みを構築した（図 2）．

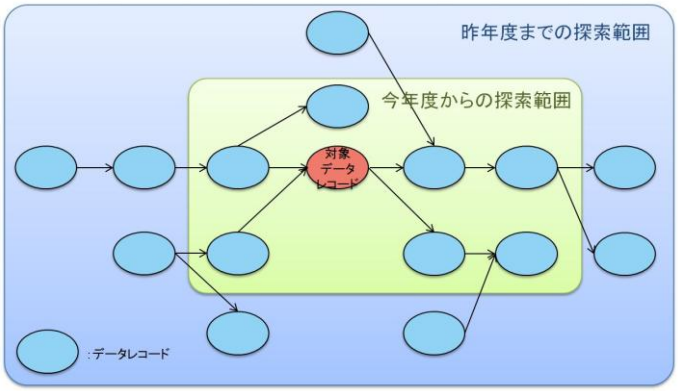

図 2　クローリングプログラムによるセマンティックウエブデータ探索範囲

また，今年度は公開に際して重点を置く特定のデータセットに対し，データ公開者の設定した多段階（現時点では最大 8 段階）のつながりを探索し、公開者がより見せたい情報，具体的には共通するデータレコードに間接的につながる同一データセット内のレコードを取得する仕組みを構築した（図 3）．これにより，ユーザーは多数のつながりデータの中から，より強いつながりを持つレコードを知ることが可能となった．

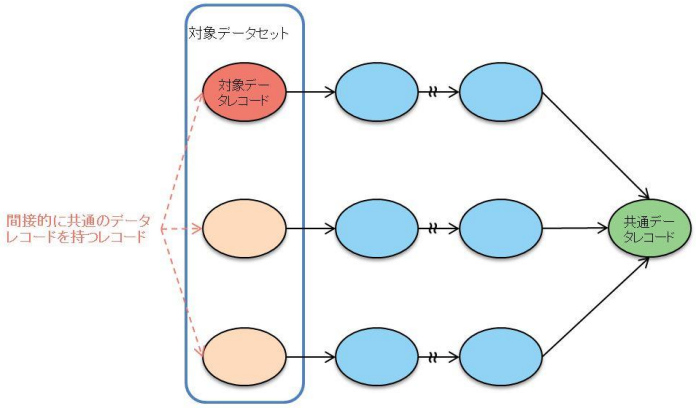

図 3　共通するデータレコードに間接的につながる同一データセット内のレコードの探索

以上の改良により，データ量とユーザーに提供する情報の質は向上したにもかかわらず，全データレコードのクローリングに要する時間は約 15 日間と，昨年度と同じ更新速度でクローリング・ジョブを運用することが可能となった．

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，理研サイネスを閲覧するユーザーに付加価値のある情報を提供し，かつクローリング時間を短縮する仕組みを実現した．今後は，プログラムの改良により処理効率をさらに向上させ，データ公開者によるデータ更新後に，より短いタイムラグで各種フォーマットの最新データを提供することを目指し，研究を進める．

6．　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．