TSUBAME 共同利用　平成28年度　学術利用　成果報告書

利用課題名　HPC を利用した自然言語処理技術の研究

英文：High Performance Computing for Natural Language Processing Technology Research

利用課題責任者
鳥澤 健太郎

所属
国立研究開発法人情報通信研究機構 ユニバーサルコミュニケーション研究所
データ駆動知能システム研究センター
http://www2.nict.go.jp/direct/

邦文抄録（300 字程度）情報通信研究機構では、Web 等に存在する大量のテキストを意味的に深く分析する技術を開発しており、WISDOM X を始めとする大規模情報分析システムを一般公開している。WISDOM X 等のシステムは、TSUBAME 2.5 環境での稼動によって、スケーラビリティと速度の向上が期待される。しかし、従来 WISDOM X を稼動させてきた計算機クラスタと、スパコンである TSUBAME の計算環境は大きく異なるため、そのままでは WISDOM X を TSUBAME 上で稼動させることはできない。そこで本課題では、WISDOM X 等で使用される高速化・高並列化ミドルウェア RaSC を、TSUBAME 2.5 環境で利用するにあたっての問題点を明らかにするとともに、TSUBAME の計算機環境で動作するようミドルウェア RaSC を拡張した。

英文抄録（100 words 程度）National Institute of Information and Communications Technology (NICT) has been developing technologies for semantic deep analysis on large-scale web texts and provided some large-scale information analysis systems including WISDOM X using the technologies. The high-performance compute nodes and the network of TSUBAME 2.5 are expected to enhance the scalability and the speed of the information analysis systems. However, the information analysis systems are incompatible with TSUBAME because they are designed for commodity hardware clusters, not for supercomputers. In this project, we aim at clarifying potential issues of using our middleware RaSC, which is an infrastructure of our information analysis systems, and extend it to work on TSUBAME.

Keywords: 自然言語処理, 大規模情報分析, テキスト分析

概要

情報通信研究機構が研究・開発している WISDOM X[1]等の大規模情報分析システムは、TSUBAME 2.5 のスパコン環境で稼動させることでスケーラビリティと速度の向上が期待できる。そこで WISDOM X 等で使用されるミドルウェア RaSC について、TSUBAME 2.5 環境への適応可能性を確認し、問題点を明らかにした。またその結果に基づいて、TSUBAME の計算機環境で動作するようミドルウェア RaSC を拡張した。

背景と目的

[1] http://wisdom-nict.jp/

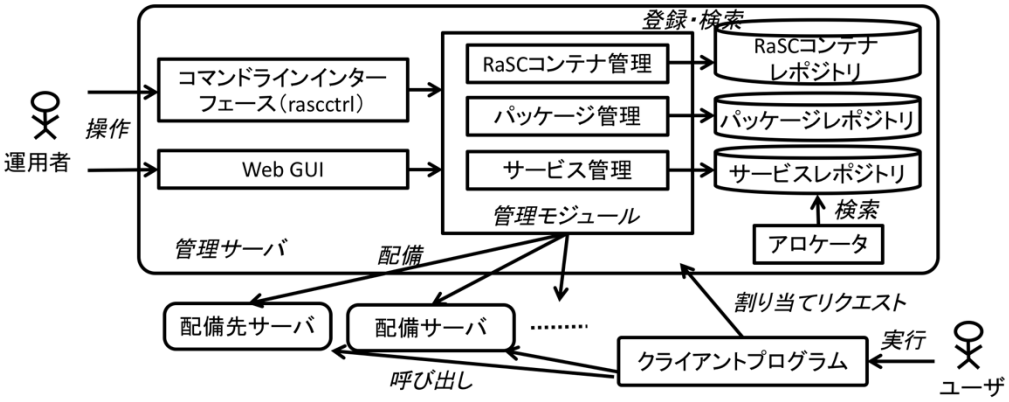

図 1 RaSC サービス管理システムの構成

情報通信研究機構 ユニバーサルコミュニケーション研究所 データ駆動知能システム研究センターでは、Web 等に存在する大量のテキストを深く意味的に分析し、情報の価値ある組み合わせや価値ある仮説を柔軟な入力を元に提示できる技術を開発している。一見かけ離れた情報間の予想もしなかった繋がりが非常に重大な帰結をもたらす事例があるなかで、情報間の組み合わせをユーザーに分かりやすい形で入手可能にすることを目指す。具体的には、文の同義性やテキストに書かれた因果関係などの事象間の意味的関係を元に、ユーザーの多様なニーズに応えられる情報やその組み合わせ、あるいは仮説を、Web 等に存在する膨大な情報源をもとに生成する技術である。情報通信研究機構では、この技術を大規模 Web 情報分析システム WISDOM X, 対災害情報分析システム DISAANA[2], 災害状況要約システム D-SUMM[3]として一般に公開している。

これらの大規模情報分析システムは、TSUBAME 2.5 環境で稼動させることにより、スケーラビリティと速度の向上が期待できる。しかし、WISDOM X 等のシステムが従来稼動していた情報通信研究機構の所有するクラスタ群と、TSUBAME の計算機環境は大きく異なる。そこで本課題では、実際に WISDOM X で使用しているミドルウェア RaSC[4]及び RaSC サービス管理システム（図 1）の TSUBAME への適応性を確認し、問題点を明らかにすることを目指した。RaSC サービス管理システムは、運用者の操作に応じて、プログラムやデータのパッケージを、RaSC コンテナと呼ぶ単位で計算機サーバに配備するものである。配備されたプログラムやデータは、RaSC の機能によって RPC サービスとしてアクセスできるようになる。WISDOM X を構成する 100 種以上のモジュールは、この RaSC サービス管理システムによって管理されている。

結果および考察

上述の RaSC サービス管理システムの TSUBAME への配備を試みたところ、以下の問題点が明らかになった。

(1) プロセスの持続的起動

TSUBAME では基本的にジョブスケジューラを介したジョブの投入としてプログラムを実行する。そのため、データやプログラムをラップする RPC サービスを持続的に起動するためのインタフェースは提供されていない。

(2) 計算ノードの指定

TSUBAME では、利用可能な計算ノードのホスト名は予約の度に変わることがある。一方、RaSC サービス管理システムでは、配備の対象となるホストの集合は固定であることが前提となっている。

(3) データ・プログラムの配置パス

RaSC サービス管理システムでは、プログラム群を計算機クラスタの各ノードのローカルディスクに配置することを前提として、クラスタごとに配備先のパスを指定するようになっている。一方、TSUBAME では共有ストレージの利用が前提となるため、ホスト名等に基づくパスの指定が必要となる。

これらの課題を解決するため、RaSC サービス管理システムの利用に先立って、タスクスケジューラに各ノ

[2] http://disaana.jp/
[3] http://disaana.jp/d-summ/
[4] https://alaginrc.nict.go.jp/rasc/

ードを確保するための持続的なタスクを投入することとした。また、TSUBAME の管理システムから取得できるノードのリストを配備先ノードとして受け付けるよう、RaSC サービス管理システムを拡張した。さらに、配備先ディレクトリのパスを、ホスト名を用いて構築できるよう、RaSC サービス管理システムを拡張した。

まとめ、今後の課題

実際に WISDOM X 等で使用しているミドルウェア RaSC の TSUBAME への配備を試み、問題点を明らかにした。また、ミドルウェア RaSC の拡張によってそれらの問題点を解決し、TSUBAME 上で動作可能とした。今後、より多くの解析プログラムやデータの配備を行い、WISDOM X 等の大規模情報分析システムについて、スパコン環境におけるスケーラビリティと速度向上を実証的に検証する。