RaSC: 深い解析を行うプログラム連携のためのミドルウェア

田仲正弘[†1] 大竹清敬[†1] 鳥澤健太郎[†1]

深い解析を行う言語処理プログラムの多くは，蓄積されたデータを処理するバッチ型であり，定常的に流入するデータに単純に適用しても実行速度が向上しない．そこで我々は，各種の解析プログラムを高速・高並列に連携可能なサービスとするミドルウェア RaSC を開発した．RaSC は，(独) 情報通信研究機構が開発した大規模 Web 情報分析システム WISDOM2013 に導入され，1 日 2000 万 Web 文書という解析速度の実現に寄与している．本稿では RaSC の概要を紹介すると共に，関連ミドルウェアとの比較を示す．

RaSC: A Middleware for Composition of Deep Analysis Programs

Masahiro Tanaka,[†1] Kiyonori Ohtake[†1]
and Kentaro Torisawa[†1]

Since most of language processing programs for deep analysis are designed for batch processing, they do not work efficiently for streaming data. Therefore we developed a middleware called RaSC, which connects a wide variety of analysis programs and enables highly scalable parallel processing. RaSC is used for a large-scale Web information analysis system WISDOM2013 and contributed to processing 20 million documents/day. In this paper, we introduce the overview of RaSC and compare the features with those of related middlewares.

1. はじめに

近年，分散処理による大規模データ分析が注目を浴びている．(独) 情報通信研究機構で開発している大規模 Web 情報分析システム WISDOM2013[2)] でも，Web から収集した数十億件の Web 文書に対して各種の深い言語処理を適用することで，質問応答を初めとする様々な分析を可能としている．

そのような言語処理プログラムは計算負荷が高い一方で，有用な分析には大量の文書の処理が必要になる．WISDOM2013 では，一日あたり約 2000 万件の Web ページを収集し，その全てに対して十数種の言語処理プログラムを適用する．しかしながら，多くの言語処理プログラムは蓄積されたデータを一括処理するバッチ型であり，日々収集される Web 文書のように，定常的に流入するデータの処理は想定されていない．そのため，大規模処理のための高速化と高並列化には，**図 1** に示したような課題の解決が必要となる．

そこで我々は，深い解析を行う言語処理プログラムを，ストリーミングでデータ交換を行うサービスとすることによって，複数の言語処理プログラムを組み合わせつつ高速・高並列で動作させることを可能にするミドルウェア RaSC (**Ra**pid **S**ervice **C**onnector) を開発した．以降では，この RaSC の概要を紹介し，関連フレームワークとの特徴について比較する．

†1 情報通信研究機構
National Institute of Information and Communications Technology

2. 高速化・高並列化ミドルウェア RaSC

RaSC は，各種の言語処理プログラムを高速化・高並列化する．以下にその主要な機構について述べる．

2.1 解析プログラムのストリーミング化

言語処理プログラムを組み合わせ，高並列で実行しても速度が向上しない主要な原因として以下がある．

- 巨大な辞書データやモデルデータのロードによる起動オーバヘッド
- 前段のプログラムの終了待ちによる，後段のプログラムの処理開始待ち

多くの解析プログラムは研究成果の実証を目的にとして作成されるものであり，その開発者たる言語処理研究者にとって，大規模処理を想定して上記の問題を解決するように開発することは困難である．そこで RaSC では，解析プログラムを高速ストリーミングに対応したサービスとする機構を提供する．この機構は**図 2** に示す構造を持つ．解析プログラムを常時起動状

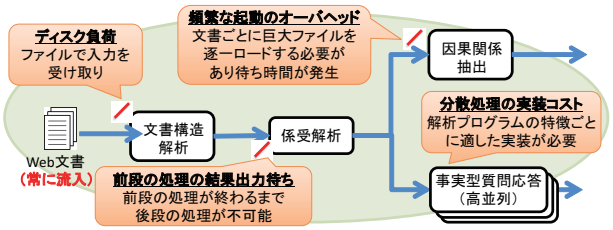

図 1　深い解析の高速・高並列実行における課題

態に保つことで起動オーバヘッドを削減すると共に，入出力を高速ストリーミング通信に対応させる．主な通信プロトコルとして MessagePack RPC を用いるが，一般的なサービス指向システムとの相互運用性を考慮し，JSON, SOAP 等のプロトコルにも対応する．

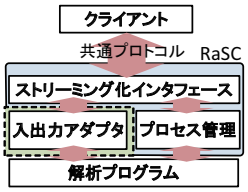

図 2　ストリーミング化機構

解析プログラムごとに数行から十数行程度のコードからなるごく小規模なアダプタを開発が必要となるが，WISDOM2013 においては，RaSC の導入によって全体での CPU 利用効率は約 2 倍に改善した．

2.2　大規模分析の並列分散化

質問応答などの高度な分析では，計算機ノード 1 台に収まらない巨大データにアクセスするため，並列処理とその処理結果の集約は必須である．しかし並列分散化に求められる処理はプログラムの性質に応じて異なり，個別に実装するのはコストが大きい．

そこで RaSC ではサービス化された解析プログラムの並列分散機構を提供する．**図 3** に示すように，利用者が定義した集約アルゴリズムの組み込み，並列実行の結果の任意のタイミングでの取得，差分データのみの転送による効率化などの機能により，様々な特徴を持つ解析プログラムに対応する．また，並列実行される各サービスと同一の RPC インターフェースを持つサービスが自動的に生成され，透過的な並列分散実行が可能となっている．

3.　関連フレームワーク

近年，Hadoop[1] を初めとする多数の大規模分散処理ミドルウェアが相次いで公開されている．**表 4** に，代表的なミドルウェアとの比較を示す．主要な違いとして，RaSC が巨大モデルデータへのランダムアクセスを行うなどの特徴を持った，深い解析を行うプログラムの高速実行を目的とすること，また研究的側面の強い非定型的な分析を構築できるよう，拡張の自由度が高いことが挙げられる．

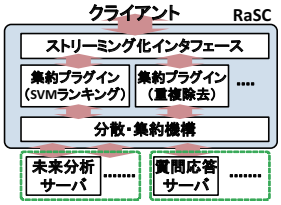

図 3　並列分散化機構

図 4　関連フレームワークの特徴

	深い解析の高速実行	大規模並列処理	簡便さ	拡張性
UIMA (NLPツール連携)	△	△	△	○
Hadoop (大規模バッチ処理)	×	◎	◎	△
Strom (大規模ストリーム処理)	×	◎	◎	△
Jubatus (分散オンライン機械学習)	△	○	◎	×
RaSC	◎	○	○	◎

4.　おわりに

本稿では，深い解析を行う言語処理プログラムを高速化・高並列化するミドルウェア RaSC を紹介した．RaSC は大規模 Web 情報分析システム WISDOM2013 の高速化に大きく寄与している．今後，公開のための整備を進めると共に，計算機環境に応じたデータの自動配置などの拡張を行う予定である．

参　考　文　献

1) Borthakur, D.: The hadoop distributed file system: Architecture and design, `http://hadoop.apache.org/common/docs/` (2007).
2) Tanaka, M., Saeger, S. D., Ohtake, K., Hashimoto, C., Hijiya, M., Fujii, H. and Torisawa, K.: WISDOM2013: A Large-scaleWeb Information Analysis System, *IJCNLP 2013 (Demonstration Track)* (2013).