ステップ1：ビッグデータ分析用テキストの抽出

佐藤 聖

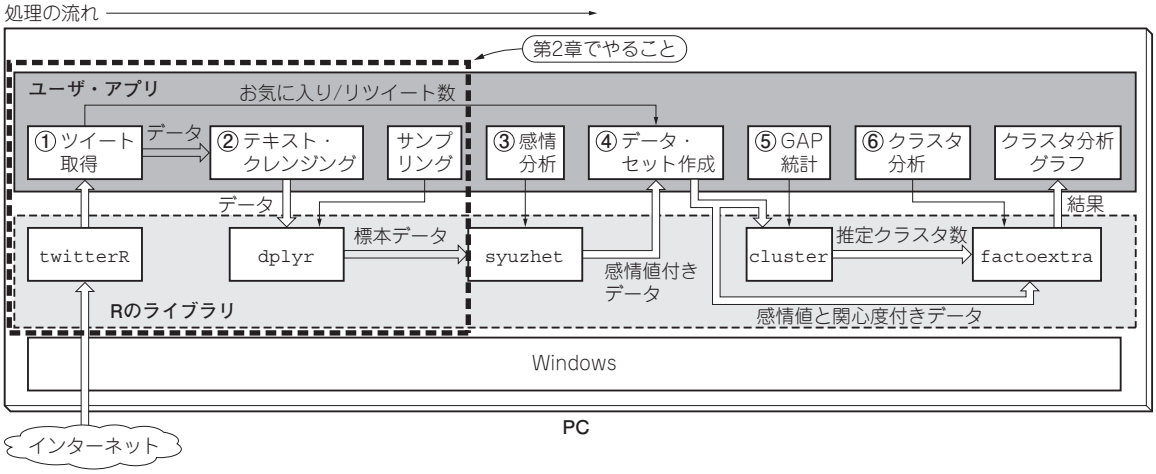

図1　第2章でやること…インターネット（Twitter）から必要なテキスト情報だけを抽出する

インターネットから集めてきたデータには，画像やHTMLのタグなど，いろいろな情報が含まれています．今回のビッグデータ解析には，これら情報は邪魔になるので，必要となるニュース（テキスト文字）だけを抽出します．本章でやることを**図1**に示します．

まずはツイートの取得

● ターゲットとなるTwitter

実験で使用するデータの配信元を以下に示します．

- ウォール・ストリート・ジャーナル・マーケット
- FXstreetニュース
- IBDインベスターズ
- Forexライブ
- USAトゥデイ
- ワシントン・ポスト

最後の2つは経済専門の新聞ではありませんが，比較対象として加えました．それぞれのタイムラインからツイートを収集しました．あらかじめRとTwitterを連携させるライブラリ「twitteR」で取得したツイートをダウンロードしました（**リスト1**）．ライブラリは仕様上の制約で最大3200件までしかダウンロードができません．

twitteRライブラリの使用準備としてTwitter Appサイトでアプリケーション登録します．処理は登録情報を用いて関数setup_twitter_oauth()でセッションの認証を行います．

タイムラインからツイート取得するため関数userTimeline()のパラメータにユーザ名と取得件数（最大3200件）を設定します．ツイートはデータ・フレームに変換後，CSVファイルを出力します．

● ビッグデータを集めるための「くふう」

データを集めるうえで想定外だったことは，1日に発信される経済ニュースの件数が少ないことでした．オンライン版の新聞記事では紙面に掲載されるメジャーなニュース，紙面に掲載されないマイナーなニュースが含まれています．

実験は各紙面から数百件から3千件のヘッドラインを収集します．多種多様なヘッドラインを収集できるはずでしたが，ツイートされているニュースはメジャーな出来事だけのようです．