オープンソースによるビッグデータ分析の可能性

足立 悠, 北島 聡

株式会社 KSK アナリティクス

haruka.adachi@ksk-anl.com, satoshi.kitajima@ksk-anl.com

概要 昨今, 産学官あらゆる業界でビッグデータ活用が進められており, データ分析ソフトウェアとして R が注目を集めている. R はオープンソースソフトウェアのため無償で利用できること, 分析パッケージが豊富に提供されていることなど多くの長所を持つが, 大規模データを扱えない, 実行速度が遅いという短所を持つため, ビッグデータ分析に適さない. 本稿では, 大規模データ分析が可能な OSS を紹介し, 実例を挙げた分析とその結果を報告する.

キーワード ビッグデータ, オープンソースソフトウェア, RapidMiner, NYSOL, Revolution R

1 はじめに

ビッグデータの登場から早 5 年が経過し, 今や様々な局面で活用が進んでいる. 例えば, 企業は購買データやソーシャルメディアデータを分析し新たなビジネスを生み出している. また, 国や地方公共団体は公共データを提供し経済活動の活性化が図ろうと試みている[1].

ビッグデータと共にデータマイニングや機械学習等の分析手法, 分析ソフトウェアも注目を集めるようになった. これまで一般に, ビッグデータのような大規模データを処理するには高価なマシン, ソフトウェアが必要と考えられていたが, 昨今のマシン処理速度の飛躍的な向上, オープンソース分析ソフトウェアの充実により, 誰もが手軽に大規模データを分析することが可能になった.

本稿では, ビッグデータ分析を支えるオープンソース分析ソフトウェアに焦点を当て, 近年の動向, 一部ソフトウェアの紹介, 実際の分析事例を報告する.

2 オープンソース分析ソフトウェア

2.1 分析ソフトウェアの動向

かつてデータ分析者は SPSS, SAS を始めとする商用ソフトウェアを利用していたが, 近年はオープンソースソフトウェアを利用する傾向が見られる. データマイニングやデータサイエンスに関する情報を提供する KDnuggets が, 2014 年 6 月に, 過去 12 ヶ月に分析プロジェクトで使用されたソフトウェアに関する調査結果を発表した[2]. その結果, 上位 10 ソフトウェアのうち半数をオープンソースソフトウェアが占めていることから, オープンソースへの期待が高まっていることが伺える.

次節以降では, 上述の調査結果 1 位の RapidMiner, 利用者が急速に増加中の Revolution Analytics R, 日本発のオープンソースソフトウェアである NYSOL について紹介する.

2.2 RapidMiner Studio

RapidMiner Studio は GUI ベースのデータマイニングソフトウェアである[3][4]. ノンプログラミングでデータ加工や分析を実施できる, 多数のビジュアル機能により様々な視点から考察できることなどの特長を持つが, 大規模データの処理には不向きである.

2.3 Revolution R Open

Revolution R は, R をより高速に, 大規模データを処理できるように拡張させたソフトウェアである[5][6]. R との 100%互換のため CRAN パッケージを利用できる, 既存の R スクリプトを利用できることなどの特長を持つが, CUI に不慣れなユーザーには扱いづらい.

2.4 NYSOL

NYSOL はデータ加工, 分析・可視化を行う複数のソフトウェアツールで構成されている[7][8]. 本プロジェクトは, 大学や研究機関などの学術界で生み出された研究成果を, 広く産業界に還元する目的で設立された.

大規模データ処理を高速に実行できる, 最先端のアルゴリズムを利用できる, ソフトウェアを無償で利用できるなどの特長を持つが, UNIX 環境での動作を前提としている(Windows 環境では, VirtualBox や Cygwin 上で動作可能).

3 オープンソースによるソーシャルメディア分析

3.1 概要

2015 年 5 月 17 日(日)に実施された大阪都構想の住民投票に関し, 投票前 40 日間の Twitter のツイートデータを収集し, オープンソース分析ソフトウェア NYSOL を用いた分析を行った[9].

3.2 分析手法

分析の流れを図 1 に示す.

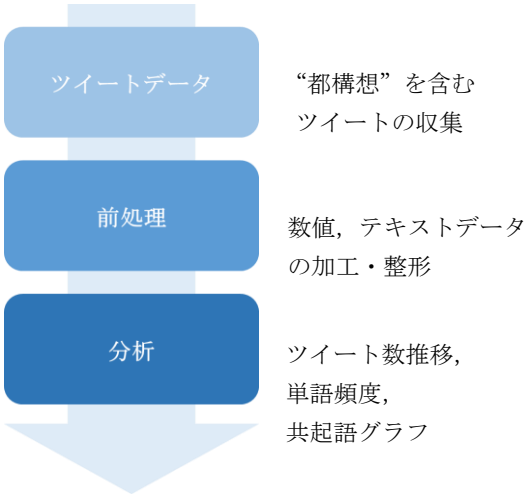

図 1 分析フロー

2015 年 4 月 1 日～5 月 10 日に，“都構想”という言葉を含むツイートを収集し分析対象とした．前処理に NYSOL の MCMD と Fumi(テキストデータのみ)コマンドを，分析に MCMD と Take，View コマンドを使用した．

3.3 結果と考察

40 日間のツイート推移を図 2 に示す.

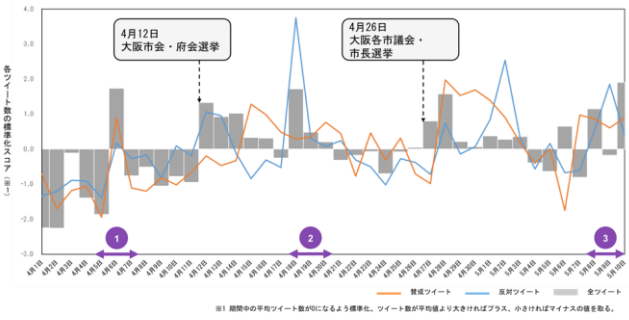

図 2 ツイート数の推移

ツイート数の増減を棒グラフで，賛成／反対ツイートを折れ線グラフで示す．賛成／反対ツイートは“賛成”または“反対”という単語が含んでいるか否かで，1 つのツイートに賛成／反対の両方の単語が含まれている場合はいずれにも含めていない(賛成／反対ツイートは，賛成／反対の単語が含むか否かを示すものであり，そのツイートが都構想について賛成／反対なのか，意味を解析して判別しているものではない)．

4 月に 2 回開催された選挙前後で相対的にツイート数が増えていることがわかる．選挙という公のイベント以外にも増減の差が大きい時期が，4 月 5 日～8 日，18 日～20 日，5 月 8 日～10 日の 3 ヶ所あり，何らかのイベントが潜んでいることが予想される．ここでは 3 ヶ所目，5 月 8 日～10 日のツイート分析結果を紹介する．

TF-IDF 法を用い，ツイートに含まれる特徴的な単語の一部を表 1 に示す.

表 1 TF-IDF 法による特徴的な単語

順位	単語	品詞
1	総統	名詞
2	断念	名詞
3	閣下	名詞
4	五輪	名詞
5	白	名詞

“総統”や“閣下”は，映画“ヒトラー最期の 12 日間”でヒトラーを橋本市長に模したパロディ動画がツイート上で拡散されたタイミングである．また 5 月 8 日に報道された，都構想が実現すれば大阪五輪もできる，といった橋下市長の発言に対するツイートも見られた．

また，単語の共起グラフを作成，クリークを列挙し，結び付きの強い単語群を可視化した．ここでは表 1 の“五輪”を含むクリークを図 3 に示す.

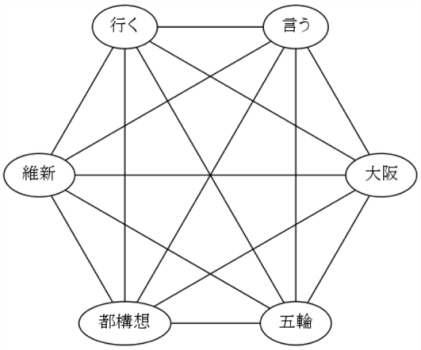

図 3 “五輪”を含むクリーク

4 おわりに

近年，オープンソースソフトウェアの機能，性能は商用ソフトウェアと比較しても遜色ないものとなっている．今後，オープンソースを使った分析を検討される際，本稿の情報を参考にしていただきたい.

参考文献

[1] 一般社団法人 オープン＆ビッグデータ活用・地方創生推進機構, http://www.vled.or.jp/
[2] KDnuggets: What Analytics, Data Mining, Data Science software/tools you used in the past 12 months for a real project Poll (Jun 2014), http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-software-used.html.
[3] RapidMiner, https://rapidminer.com/
[4] RapidMiner (パートナー), http://www.rapidminer.jp/
[5] Revolution Analytics, http://www.revolutionanalytics.com/
[6] Revolution Analytics (パートナー), http://www.revolutionanalytics.com/
[7] NYSOL, http://www.nysol.jp/.
[8] NYSOL.biz (パートナー), http://www.nysol.biz/
[9] プレスリリース：大阪都構想に関する 40 日間のホットワードが判明、ツイートの声を可視化, http://ksk-anl.com/wp-content/uploads/2015/05/PressRelease201505.pdf