不揃いなデータ達の分析を行う前の Tips

榊剛史 [1], 吉田光男 [2], 伊川洋平 [3], 佐々木一 [4], 杉原太郎 [5]

1 ホットリンク, 東京大学 2 豊橋技術科学大学 情報・知能工学系

3 IBM 東京基礎研究所 4 東京大学策ビジョン研究センター

5 岡山大学大学院自然科学研究科

1 はじめに

Web が登場して二十余年が経過し, Web を通じて大規模かつ多様なデータにアクセスすることが可能となった. 2013 年末時点で, Web 上には約 7 億 6 千万の Web サイト, 14 兆以上の Web ページ, 672 エクサバイトのデータが存在しているとの報告もある[1]. このような現状を受け, Web 上のデータを用いて様々な研究やビジネスを行う流れが加速している. オープンデータ運動や IoT (Internet Of Things)も普及しつつあり, Web 上のデータを活用するという流れはとどまらないであろう.

Web 上のデータが持つ特色の一つは, その多様性である. 目的に合わせて多種多様な表現ができるという柔軟性の高さが Web の特徴であり, 普及した要因の一つであると考えられる. 一方, 柔軟性が高いという特徴を裏返せば. それらのデータが不揃いであると言える. 多種多様なデータが Web 上には存在するが, それを自らが知りたいことのために用いようとすれば, それらのデータを収集・蓄積し, 目的に合わせて整形・加工する作業―前処理―が必要不可欠となる.

ところが, その重要性にも関わらず, データの前処理について知見を共有し, 議論される場は非常に少ない. これは, データ分析, Web マイニングという言葉の持つ先進的なイメージに対し, データの前処理のために必要となる作業は泥臭く, 地道なものとなるためであると推測される. 学術的にもこれらの前処理に学術的な新規性が認められる可能性は低いため, 論文や発表においてはデータの前処理は数行の記述, 数分の説明に集約されてしまう. 結果として, これらの前処理に関する知見やノウハウを身につけるための障壁が高く, Web マイニングへの参入障壁になっていると考えられる.

本企画では, Web 上のデータの前処理に関するノウハウを, 学術およびジビネスの分析専門家から紹介いただき, 分野全体を前進させるために共有することを目指す. 同時に, これら前処理のノウハウや知見を公開・共有するためにはどのような仕組みや仕掛けが必要であるかについて, 議論を通じて明らかにしていきたい.

2 登壇者

2.1 吉田光男

Web マイニングを行うには, Web 上から様々な情報をクロールし, それを分析しやすい形で蓄積する必要がある. そこで, Web 上の情報のクロール・蓄積について気をつけるべき点・工夫すべき点について紹介する.

2.2 伊川洋平

ソーシャルメディア分析は位置情報を扱うことで, 分析の幅が大きく拡がる. 本講演では, ソーシャルメディアとそれに付与された位置情報を扱う上で注意・考慮すべき点について説明する.

2.3 佐々木一

Web データに留まらず多様なデータを扱えることは分析の幅を広げる一方で, 不慣れなデータの前処理がプロセス上のボトルネックとなる。前処理の位置付けについて認識を共有し多角的に議論したい.

3 進行方法

ディスカッションは, 各登壇者が研究を進めてきた中で経験した, 成功や失敗, 学会・研究会での質疑応答などを発表し, 短く質疑応答行う. その後, 本チュートリアルで議論したいポイントを司会が提示し, 登壇者同士, および会場との質疑応答という流れで行う. 本セッションにおいては, 登壇者のみならず会場の参加者の方々からも積極的に情報共有及び議論への参加をお願いしたい.

[1] http://www.factshunt.com/2014/01/total-number-of-websites-size-of.html