2014年総会シンポジウム「ビッグデータの可能性と課題――監視・シミュレーション・プライバシー」・論文

社会科学におけるテキスト型BIG DATAの可能性

Impact of Text-based BIG DATA on Social Sciences

キーワード：
テキスト型BIG DATA，社会調査，社会シミュレーション
keyword：
Text-based BIG DATA, social research, social simulation

芝浦工業大学システム理工学部　中　井　　豊
Shibaura Institute of Technology　Yutaka NAKAI

要　約
BIG DATAと言えば，個々人の行動に関する定量的なDATAに注目が集まるが，意味を専ら取り扱う社会科学では，テキスト型のBIG DATAの利用が期待される。「生」の，「本音」の，「時々刻々」の声を把握できるテキスト型のBIG DATAによって，社会の通念等を，個人レベルの分解能で常時観測できることで，社会調査のあり方が根本的に変わる可能性がある。また，社会理論研究に関しては，従来，理論の妥当性を経験的に検証することは難しかったが，社会シミュレーション等他の技法とともに，テキスト型BIG DATAの利用が進み，検証過程の説得性を飛躍的に高めるであろう。一方，膨大なデータの中から如何に意味を自動抽出するかが大きな課題である。形態素解析や共起解析などの手法が開発されているが，自然言語処理における一層の技術開発が望まれる。

Abstract
Text-based BIG DATA seems highly hopeful in social sciences like sociology dealing with a meaning. The usage of BIG DATA will make us capable of observing our society. Because the data gives us a collective of individual live and real time thinking, it will give much effect on empirical researches drastically. BIG DATA will also be used in theoretical researches. Although it has been difficult to verify social theory, BIG DATA will enhance the quality of verification on social theory together with the usage of social simulation etc. On the other hand, the state of arts of analysis on text data is not enough. The improvement is highly expected in social sciences.

2014年総会シンポジウム「ビッグデータの可能性と課題――監視・シュミレーション・プライバシー」・論文

社会科学におけるテキスト型BIG DATAの可能性

Impact of Text-based BIG DATA on Social Sciences

キーワード：
テキスト型BIG DATA，社会調査，社会シミュレーション
keyword：
Text-based BIG DATA, social research, social simulation

芝浦工業大学システム理工学部　中　井　　豊
Shibaura Institute of Technology　Yutaka NAKAI

要　約
BIG DATAと言えば，個々人の行動に関する定量的なDATAに注目が集まるが，意味を専ら取り扱う社会科学では，テキスト型のBIG DATAの利用が期待される。「生」の，「本音」の，「時々刻々」の声を把握できるテキスト型のBIG DATAによって，社会の通念等を，個人レベルの分解能で常時観測できることで，社会調査のあり方が根本的に変わる可能性がある。また，社会理論研究に関しては，従来，理論の妥当性を経験的に検証することは難しかったが，社会シミュレーション等他の技法とともに，テキスト型BIG DATAの利用が進み，検証過程の説得性を飛躍的に高めるであろう。一方，膨大なデータの中から如何に意味を自動抽出するかが大きな課題である。形態素解析や共起解析などの手法が開発されているが，自然言語処理における一層の技術開発が望まれる。

Abstract
Text-based BIG DATA seems highly hopeful in social sciences like sociology dealing with a meaning. The usage of BIG DATA will make us capable of observing our society. Because the data gives us a collective of individual live and real time thinking, it will give much effect on empirical researches drastically. BIG DATA will also be used in theoretical researches. Although it has been difficult to verify social theory, BIG DATA will enhance the quality of verification on social theory together with the usage of social simulation etc. On the other hand, the state of arts of analysis on text data is not enough. The improvement is highly expected in social sciences.

1　社会科学とBIG DATA

BIG DATAには大きな期待が寄せられている。個々人の時々刻々の行動，特に，消費（購買・決済データ）や移動（位置データ）をありのままを把握できる点は画期的であって，経営学，経済学，社会工学等で大きな期待が寄せられている。例えば，経済学では，精緻な一般均衡モデルによる経済予測が既に一般的であるが，株価の変動をBIG DATAとして解析しクラッシュを予測しようという研究が積み重ねられているし，社会工学では，交通渋滞や災害避難誘導等で，個々のエージェントに判断を委ねた複雑系のシミュレーションモデルが開発されつつあり，BIG DATAを使って，シミュレーションのリアリティを飛躍的に高めようとしている。

一方，政治学や社会学といった分野では，これらの学問が，思想・信条・選好といった意味の取り扱いに注力するという特徴によって，経済学等と比べ，BIG DATAに対する期待感に温度差がある。また，理論研究においてシミュレーションの利用等が始まったばかりであり，例えば，理論社会学では抽象的な一般交換モデルが専ら探求されており，信頼や安心など社会を支える基礎的なメカニズムの定性的な理解に供されている段階であって，BIG DATAの利用は進んでいない。この様に，同じ社会を対象にする科学でもBIG DATAに対する距離感が違う。つまりは，BIG DATAは個々人の行動に関する「量」的データの集積として期待されている。一方，新聞，雑誌，TV等マスメディアには膨大な言説（世論・通説・常識）が流通し，インターネットではSNSを通じて個人の思想・信条・選好の表出が爆発的に広がっている。これらは質的なテキストデータであるとともに，正にBIG DATAそのものに相違ない。そして，意味の取り扱いを生業とする社会科学だからこそ，今後益々，テキスト型BIG DATAに対する注目が集まるものと思われる。以下，具体的に，そのポテンシャルを，実証研究と理論研究の両者に与えるインパクトの面から整理してみよう。

2　常時の社会調査の可能性

経験科学である社会科学で調査は極めて重要である。アンケート等の社会調査を通じて，（事実としての行動に加えて），個々人の思想・信条・選好を観測し，これを基に，実証研究が進められてきた。ところで，取得された思想・信条・選好は，あくまで調査時点のデータに過ぎず，調査が行われた正にその時点の世論等に強く影響されてしまう。また，回答者が調査紙（画面）に対峙して回答するという調査環境が，個人の考えをありのままに引き出しているか定かでない（例えば，選好の調査を考えて欲しい。机に座ってじっくり頭で考えて出てくるものが本当の声と言えるだろうか。選好とはその場に直面して初めて，自分の中に潜む選好に気付くものではないか）。更に，原理的な限界として，如何に周到に準備しても，アンケートには想定外の回答というリスクが存在する。質問者が想定していたフレームとは異なるフレームで人々は思考していたという事態である。周知の通り，アンケートは「仮説の検証」に適した研究手法である。従って，仮説Aが棄却されたら，新たな仮説Bを立て再調査しなければならず，運が悪ければこれが延々と続く。 この様に，従来の社会調査には，様々な原理的限界が存在する。そしてこの意味で，「個人レベル」の，「生」の，「本音」の，「時々刻々」の声を把握できるテキスト型のBIG DATAは，社会を経験的に科学しようとする者にとっては，理想的なデータと言えよう。言い換えれば，通説・常識や信条・選好をその変動も含めて常時観察できる様になる訳で，ハッブル宇宙望遠鏡が天文学にもたらした革命にも似たインパクトが，社会調査においても期待される。では，アンケート調査が要らなくなるのだろうか？言うまでもなく，アンケートは仮

説の検証を行うもので，この点，無くてはならない研究手法として利用され続けるであろう。ただし，仮説を準備しなければならない。前述の通り，準備した仮説が見当違いなものであると調査は失敗する。もちろん，このリスクを低減するために，事前インタビューなどを併用されるわけであるが，もし，膨大なテキスト型BIG DATAの意味解析によって仮説の構成支援されれば，リスクを抑えた質の高い調査が可能となるであろう。つまり，BIG DATAから仮説を抽出し，社会調査による「仮説検証」につなげる形で，互いに補完し合うことになろう。 この意味で，テキスト型のBIG DATAは，意味を扱う社会科学の実証作業を一段飛躍させうるポテンシャルを持つ。

3　社会理論検証の可能性－仮想エルサレムを事例に

次に，社会理論とテキスト型BIG DATAの関係を考えよう。そして，社会理論研究の１手法として社会シミュレーションを想定しよう。言うまでもなく，シミュレーションは経済学や経営工学の分野で定量的な予測手法として盛んに利用されているが，近年，社会学や政治学（国際関係論を中心に）においても利用が広がりつつある。これらの分野のシミュレーションでは，抽象的なアルゴリズムで意味を表現し，意味と因果の発現を結び付け，社会・政治現象への理解を深めようとするもので，質的なシミュレーションとも言えるが，理論研究に共通の課題として，モデル（と結果）の妥当性を示すことが難しい。ここで，先進事例を挙げて，妥当性の検証の仕方を検討してみよう（なお，本事例は，BIG DATAを利用した事例ではないことに注意しよう）。

Bhavnani等（2011）は，エージェント・ベース・シミュレーションを用いて，エルサレムの民族紛争モデルを構成した。 エルサレムでは，よく知られる様に，パレスチナ人，正統派ユダヤ教徒，超正統派ユダヤ教徒間で民族紛争が続く（図Ａは人口構成を示す）。紛争の統計が地区毎に時系列で蓄積されており，例えば，図Ｂは2001-2009年の間の累積紛争回数の分布データである（濃い部分で紛争が多い）。 Bhavnani等は各民族間の社会的距離という概念を使って，民族ｉとｊの距離が閾値を超えれば超えるほど紛争発生の確率が高まるロジット型の関数を導入する。社会的距離とは，各エス二シティ間の，文化的，政治的，経済的，社会的な相違による緊張の度合いを表現するものと説明され，また，閾値は，各民族ｉが認知する被差別の度合い，各地域で起きた過去の紛争の記憶，各地域に投入される治安力によって変動する。そして，仮想エルサレムの人口構成を実際と同じにして，紛争の発生をシミュレーションする。

具体的には，①2001-2005年の紛争データを訓練データとして用い，訓練データとシミュレーションの食い違いをミニマムにする様な，つまり訓練データを最もよく説明し得る様な，各民族間の社会的距離と各民族 ｉが認知する被差別の度合いを探索する。そして，②この作業とは独立して，パレスチナ社会で流通する言説（新聞や放送）からその時点での社会的距離や被差別の程度を，定性的に判断し，通説とする。そして最後に，③探索した社会的距離や被差別度と通説が同型であることを確認する。更に，④構成したモデルを使って，2006-2009年の紛争発生分布を予想し（図Ｃ），現実に起こった紛争と比較して，彼らのモデルが妥当であると主張する。（図Ｄがその結果である。予測と現実が一致するほど図が白くなるが，確かに一致を確認することが出来る。）この様に，Bhavnani等の研究は，エージェント・ベース・モデルによって民族紛争をシミュレーションし，社会統計データ（BIG DATAではない）を用いて定量的な検証を試みた数少ない研究である。

ところで，妥当性の検証過程を振り返ると，④の過程は説得力があるが，③の過程は心もとない。

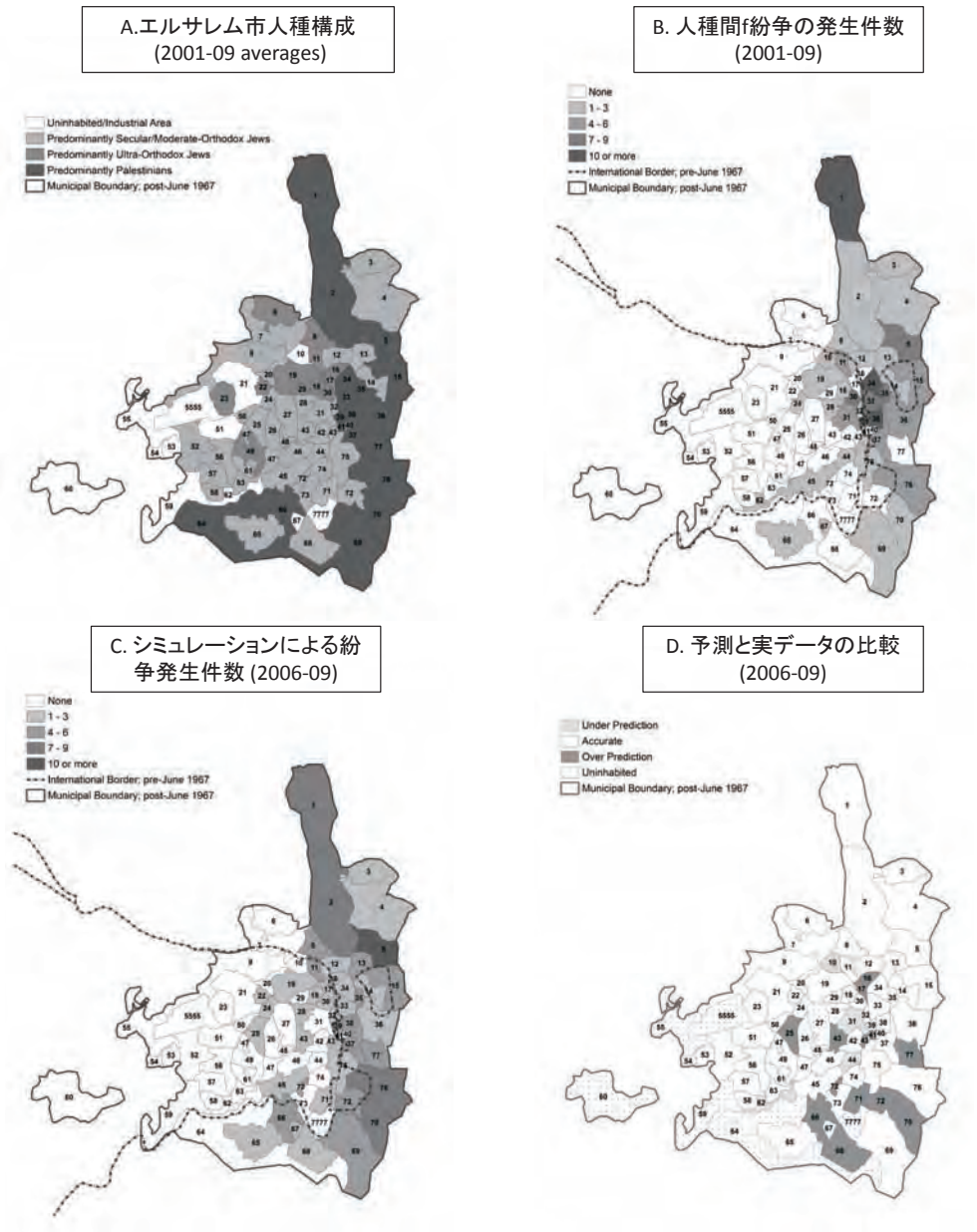
西欧や中東の研究者であれば差別の度合いなどは肌感覚で理解できるのかもしれないが，例えば，極東の人間である我々には難しい。そして，この部分に，テキスト型BIG DATAの分析が貢献する余地があると思われる。「民族i が嫌い」だとか「我々は差別されている」といった信条は，ネット上の言説に溢れている。SNSやblogなどで表明されるこの種の言説をまるごとBIG DATAとして収集し，社会的距離や被差別の程度を抽出することが出来れば，モデルの説得力は格段に高まろう。そして，妥当性が了解されれば，紛争解決のための実践的な政策提言に一層のリアリティが加わるものと期待される（実際，彼らは政策提言を志し，1967年境界線を再設定した場合の効果を示している）。

4　意味解析の現状と期待

以上の様に，実証と理論研究双方に対して，大きな可能性を秘めるテキスト型BIG DATAであるが，大きな壁が立ちはだかっている。テキストデータの自動解析である。現在，クローラーやTwitterbotなどのプログラムを使って，SNS上の膨大なテキストデータを自動的に収集，アーカイブ化することが可能となっている（例えば，国立国会図書館のインターネット資料収集保存事業が有名である）。次に，集めたテキストデータをテキスト分析にかけることになるが，これには，①形態素解析（単語に分けそれぞれに品詞を付与すること），②構文解析（単語間の構造を同定すること），③意味解析（文章の意味を同定すること），の３段階がある。そして，現時点の技術水準は概ね，①の実用段階にある（有名な形態素解析器には例えば茶筌（ChaSen）やMeCabがある）。一方，最大の壁は，好き・嫌いの程度の判定といった③の意味解析の自動化であるとされる。ただし，意味抽出の萌芽段階の手法であれば，テキストマイニング技術が広がりつつある。少し具体的に説明しよう。まず，テキストマイニングで一般的に行われているのは，共起解析である。単語Aと単語Bが同一文章上に使用された時，単語Aと単語Bは共起したという。この考え方を基に，分析対象とするテキストの中の共起関係を全て調べ上げ，諸単語間の共起関係を行列で表すのが共起分析である（有名なテキストマイニングのツールにはKH corderがある）。そしてこの共起行列を多変量解析にかけることで，単語間のクラスター構造が共起ネットワークとして可視化され，当該テキストで何が語られていたのかが浮かび上がり，第３者に提示可能な客観的なデータとして出力される。ただし，この作業は容易ではない。特にネット上の言説には，タグや絵文字など様々なノイズが混在しており，これをクリーニングしなければならない。当然，手作業では不可能であるが，ノイズのパターンは不規則であり，クリーニングには多くのノウハウが必要で，自動化への道のりは遠い。また，共起ネットワークは単語間の距離を示してくれるので，そのテキストが扱うアジェンダが可視化される訳だが，例えば，当該テキストが否定的か肯定的かまでを自動判定するのは難しい（二重否定からニュアンスまでを考えると難しさが想像できよう）。更に，共起ネットワークがアジェンダを示してくれるとは言え，図が何を物語っているかは，当該テーマの専門家でなければ読み解けない（言い換えれば，今後とも専門家の解釈は必要であり続けるということである）。以上が意味解析の現状であり，一層の技術開発の進展が待たれる次第である。

自然言語処理技術の進歩は目覚ましく，意味解析で出来る範囲は確実に広がってゆく。結局，社会学や政治学の研究が，どこかで思想・信条・選好に関わらざるを得ない以上，テキスト型BIG DATAの利用は確実に広がってゆく。筆者の専門である社会理論のシミュレーションで言えば，近い将来，ネット上の膨大な言説から，思想･信条･

選好を抽出し，これを基に個々のミクロなエージェントを構成し，次に多数のエージェントを相互作用させて現実のマクロな社会現象を再現し，社会統計やBIG DATAから抽出した通説等と比較検証する，といった研究が増えてくるであろう。質的なBIG DATAは，高度に抽象的な社会理論研究においても，検証過程の説得性を飛躍的に高めるものと期待される。



図－１　仮想エルサレムによる民族間紛争のシミュレーション（Bhavnani et al 2011）

参考文献

国立国会図書館インターネット資料収集保存事業 HP, <http://warp.da.ndl.go.jp/>

奈良先端科学技術*大学院*大学 松本*裕治*研究室 HP, <http://cl.naist.jp/>

MeCab公　式Webサ　イ　ト, <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>

Bhavnani, Miodownik, Donnay, Mor, and Helbing (2011) Residential Segregation and Violence, Presentation given at the conference CCSS 2011, ETH Zurich.

KH Cordcr HP, <http://khc.sourceforge.net/>