

ソーシャル e サイエンスのための コンテンツ作成支援技術

独立行政法人産業技術総合研究所 社会知能技術研究ラボ長 橋田 浩一

PROFILE

1986 年東京大学大学院理学系研究科博士課程修了。理学博士。1986 年電子技術総合研究所入所。1988～1992 年(財)新世代コンピュータ技術開発機構に出向。2010 年 2 月より産業技術総合研究所社会知能技術研究ラボ長。セマンティックコンピューティングおよびその応用としての文脈依存型サービス、知の社会的共創等の研究開発に従事。所属 社会知能技術研究ラボ長。2008 年度から特許産業日本語委員会委員長。



複雑化し多様化する諸問題に対処し、社会の持続可能性を担保するには、社会全体の知識創造能力を維持し向上させる必要がある。それには、日常生活や業務においてデータを蓄積し活用しつつ多様な仮説を構築・反証・改良し続ける知識創造活動を社会の隅々まで広げることが望ましい。これは、生活や産業で定常的に生み出されるデータの構造化・連携・共有・分析とそれにまつわる協働等の社会全体の知的機能を高度化するための、人間と調和した情報技術を必要とする。こうして情報技術により社会の日常の営みと融合した科学・技術の研究・開発をソーシャル e サイエンス (social e-science) と呼ぶ。e サイエンスとはデータや装置を情報ネットワークで共有して協働することによる科学研究だが、ソーシャル e サイエンスはそれを社会の通常の機能と融合したものである。

仮説検証サイクルによる業務改善 (いわゆる PDCA サイクルなど) の考え方はすでに普及しつつある。しかし、多数のステークホルダが参画し市場全体を巡るような大きなサイクルが健全に機能しているとは言い難い。また、仮説の構築やデータの蓄積・分析において情報技術などを活用する余地もまだ大きい。

社会と科学とを融合するには、特に、生活や業務における価値と科学研究 (仮説の構築とデータの分析) における価値をともに高めるようなデータのライフサイクル (生成・利用・蓄積・分析・再利用) を設計することがきわめて重要な課題だろう。それには、データを観測・作成時に分析することにより生活や業務の意味に応じて

構造化したコンテンツにしておくのが有効と考えられる。運動データやバイタルデータを分析して人間の行動や状況を表わすコンテンツとし、購買や医療や移動などの多様なサービスの利用履歴と併せて相互関係を明示するように構造化することにより、それらのサービスを利用者主導で連携させたり、履歴データを統合的に分析して生活を最適化したりできるだろう。たとえば、病院やレストランで体調やスケジュールに合った治療を受けたり食事を取ったりすることが考えられる。また、テキストデータを分析して意味的な構造や語義を明示することにより、翻訳や検索や要約などの自然言語処理技術に基づくサービスや統計的分析の品質を向上させることができる。さらに、そのように構造化された自然言語のコンテンツを用いて自動的にさまざまなサービス (たとえばホテルの予約やオンラインショッピングなど) を利用したりすることも考えられる。

以上のように構造化によってデータの利用や分析における付加価値を高めるには、構造化の範囲を広げ精度を高める必要がある。範囲を拡大するには、同種のデータの間でデータ形式等の差異を解消する何らかの正規化が必要である。構造の精度を高めるには、付加的な情報を与える必要がある。

正規化によって同種のデータの間での形式等の差異を解消するというのは、たとえば、ダイナミックレンジや精度が異なる複数のセンサがあるときに (もちろん目的に応じて) その差を軽減するとか、複数の病院の電子カルテを同じ形式に揃えとかいったことである。データ

形式については標準規格があればそれに合わせて正規化すれば良いが、実際には技術とその利用形態の進歩に標準化が追い付かないことが多いので、従来の標準化の手続きよりもさらに迅速な手法が求められる。本稿では詳細に立ち入らないが、それは、共通のオントロジーを集合知によって拡張する方法と、個別のスキーマとこの共通オントロジーの間の変換スクリプトを同じく社会的に共有し拡張する方法を含むことになるだろう。

一方、構造の精度を高めるための付加的な情報は当然ながら人為的に与えるものである。人間の負担を小さくするため、その付加情報は人間の行動によってデータが生成される機会を捉えて生成と同時に与えるのが望ましい。たとえば、携帯電話を切符として駅の改札を通ったり屋内のBluetooth端末と通信したりすることによって人間の正確な位置が判明すれば、その前後の運動データは徒歩の様子を表現していることがわかったりするだろう。

付加情報を与える際に人間が何らかの作業をせねばならないこともある。一般には、データの生成時にそれを解析してその結果を人間にフィードバックし、これに対する人間の反応から付加情報を得るということになるだろう。典型的な例として、自然言語のテキストを入力している際にそのテキストを自動解析し、入力の修正や補完をしつつ、正しく解析されたテキストデータを作成することが考えられる。修正や補完はその候補を人間に示して適切なものを選択させることになるだろう。従来は入力完了後に解析を行っていたが、これを入力時に行ない、その際の人間とのインタラクションによって構造の精度を向上させるわけである。

ここで入力されるテキストは産業日本語のように何らかの形で制限されたものになると考えられる。そこで利便性を損なわないようにするには、表層の表現の形は制限しても、表現される意味内容は制限しないことが重要である。また、コンテンツ作成の生産性を向上させる(少なくとも低下させない)ためには、入力の修正や補完によって入力作業の速度が高まる(あまり低くならない)必要がある。

一般的な文書作成にこの方法をいきなり適用するのは困難なので、たとえば電子カルテの特定の入力スロットなど、意味内容の範囲が限定されたテキストの入力に適用し、その範囲の限定を利用して入力の修正や補完の精度を高めることによって入力作業の効率向上を図るのが適切だろう。そのようなテキスト入力においては、初心者にとっては入力支援機能の有用性が高く、熟練者はその機能をあまり気にせずに制限された表現をすらすら入力できる、というのが目指すべき利用形態だろう。

やや技術的に細かい話になるが、入力支援(特に入力補完)においては、一般化LR統語解析(generalized LR parsing)のような方法が有効のように思われるかも知れない。しかし、一般化LR法では意味的な予測ができない(より正確に言えば、意味的な制約をLR表に埋め込む技術が確立していない)ので、むしろ左隅統語解析(left-corner parsing)のような方法を用いるのが適切と考えられる。左隅統語解析において、未完成の各統語範疇の左隅の統語範疇の候補を意味構造込みで有限個に限定することにより、解析中のデータを効率的に管理することができる。

このような技術を確立するには、入力支援そのものに加えて、辞書と文法とオントロジー(意味構造のスキーマ)を日常的に保守拡張する技術の開発が必要である。そのための研究開発にもソーシャルサイエンスのような方法が相応しい。