テキストマイニングを利用した感染症監視システム

ナイジェル・コリアー　Nigel Collier
総合研究大学院大学准教授　情報学専攻／情報・システム研究機構　国立情報学研究所准教授

川添　愛
津田塾大学　特任准教授

**感染症を素早く発見し適切に追跡するための手段として、
世界ではウェブ情報のテキストマイニングによる自動監視システムが開発されている。
この「情報学と公衆衛生の連携」の前には大きな可能性が広がっている一方で、解決すべき課題も多い。**

公衆衛生および医療の専門家にとって、ウェブ上を流れる情報は、人びとの健康を脅かす出来事を地球規模でタイムリーに把握するための手がかりとして、大いに期待できるものである。しかしウェブの情報は雑多で整理されていないうえ、誤解を招きやすく、言語もさまざまであるため、効率的な利用には情報処理技術によるサポートが不可欠である。ここでは、テキストマイニング技術の応用によりウェブから有用な事実を引き出す試みを紹介する。

情報学と公衆衛生の連携

過去10年間に、感染症などの緊急事態への対応において、ウェブベースの自動監視システムが効果を上げている。それらの多くは、ほぼリアルタイムに配信される豊富なオンラインニュースを情報源とし、さまざまな文体や表現で記述されたテキストから、テキストマイニングにより必要な情報を取り出し、状況の把握に役立てている。

このような「情報学と公衆衛生の連携」は、1990年代半ばにカナダ公衆衛生局（PHAC）がスタートさせた世界公衆衛生情報ネットワーク（GPHIN）に端を発する。このシステムは世界保健機関（WHO）と連携し、2002年後半には重症急性呼吸器不全症候群（SARS）の発生を検知したことで知られる。

ほぼ同じ頃、国際感染症学会（ISID）がProMED-mailという情報ネットワークを立ち上げている。これは、専門家がボランティアで疾病の発生を報告するもので、現在165カ国から約4万人が参加している。このシステムは、自動的なシステムによる監視の有効性を確かめるうえでの一つの基準となっている。

GPHIN以外に国家が助成しているシステムには、欧州委員会（EC）のMedISysや、米国ジョージタウン大学のProject Argusなどがある。研究者向けのシステムには、HealthMap、PULS、EpiSpider、Google Flu Trend、そして、われわれが開発しているBioCaster（バイオキャスター）がある。これらの多くが2009年春の新型インフルエンザ（H1N1）の早期発見に成功し、地球規模での追跡に寄与したと報告している。

事実を掘り起こす

テキストマイニングは、自然言語処理の長年の研究成果に基づいている。この技術の主なプロセスは、さまざまな形式のテキストを収集・整理し、重要な概念を表す用語（人名、地名、組織名、病名など）とその同義語を検知し、関連するものを結びつけるというものである。結果として、構造化されたデータベースに利用できる形式で、事実を取り出すことができる。

テキストマイニングを行うには、コンピューターに十分な専門知識を与える必要がある。コンピューターウイルスと病原ウイルス、また過去の出来事と現在の出来事を混同するようなミスを防がなくてはならないからである。専門知識の大部分は、SRL（Simple Rule Language、http://code.google.com/p/srl-editor/）のような規則ベースの専門エンジンや、BioCaster Public Health Ontology（http://biocaster.nii.ac.jp）のようなオントロジーを用いて記述される。

技術的には多くの課題がある。まずは重複検出の問題で、トピックを認識して複数のテキスト間の内容の重複を発見する必要がある。特に感染症の発生直後は、「トロント在住の35歳の男性がSARSに感染」と「オンタリオ州東部に住む2児の父が重い呼吸器不全を発症」のように、共通のキーワードがない重複が多く見られる傾向があり、検出にはより深い意味理解が要求される。また別の課題として、地名など、文脈に強く依存する概念の理解がある。たとえばCamdenという地名は、Wikipediaによれば5カ国に20以上も存在している。さらに、情報の信頼性や真偽の判別は、長期的に解決すべき問題である。

傾向を見つけだす

一般に、リスク評価を行うには「どこが異常なのか」を理解する必要がある。熟練のアナリストは、時系列的な傾向や因果関係モデルに直感的な洞察を組み合わせて異常を認識する。浅いレベルの意味理解を可能にするテキストマイニング技術は、人間のアナリストに取って代わるというよりもむしろ、彼らを手助けして時間の有効活用に貢献するものである。

現在、われわれが独立行政法人科学技術振興機構（JST）さきがけプログラムの助成により推進している研究は、感染症の報告から「異常さ」を示す特徴を発見することを目的としている。疫学には異常を検知するための精緻な統計モデルが存在するが、われわれはこれを言語データにも適用し、感染症の危険性をよりタイムリーに予測することをめざしている。

傾向を見極めるうえでは、メディア報道におけるバイアスが障害になることがある。たとえば、国際的なメディアと各国の国内メディアでは、注目する出来事や報道の仕方に大きな違いがある。またデータの可視化も重要であり、頻度分布、地理的な分布、時系列分析、外的リソースへのリンクなど、専門家の知識発見を促す情報の効果的な提示は、今後取り組んでいくべき重要課題である。

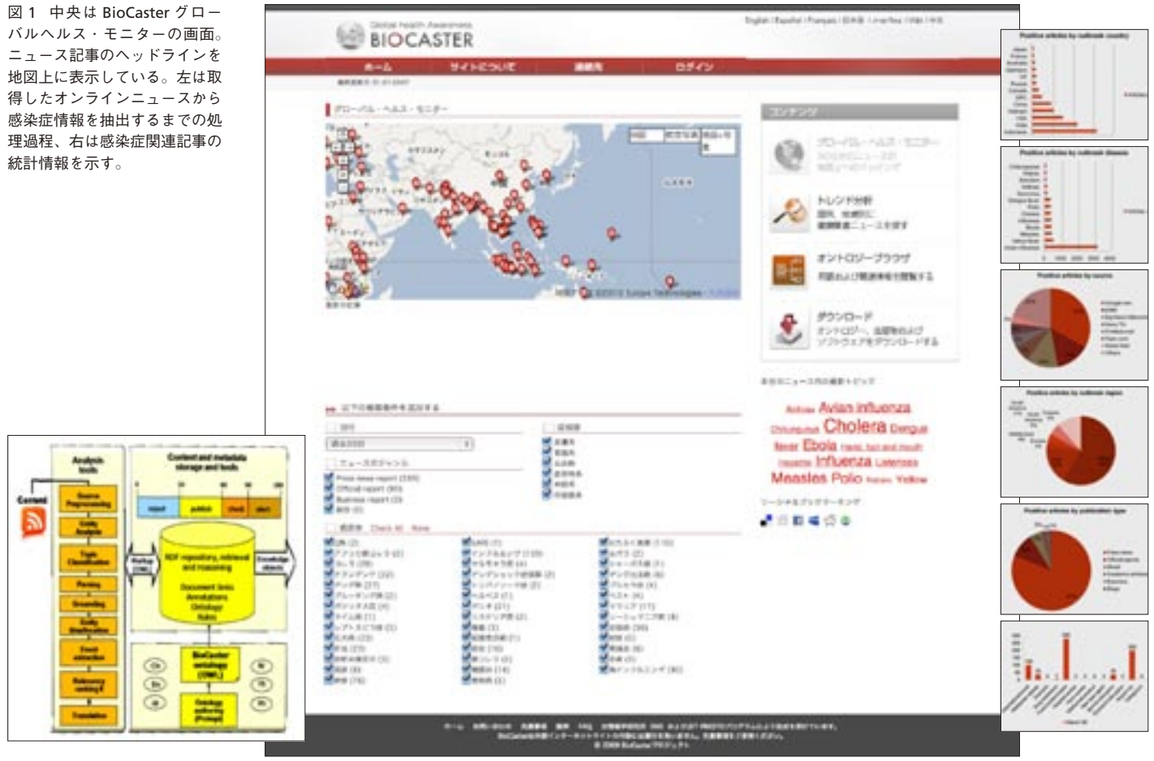

図1　中央はBioCasterグローバルヘルス・モニターの画面。ニュース記事のヘッドラインを地図上に表示している。左は取得したオンラインニュースから感染症情報を抽出するまでの処理過程、右は感染症関連記事の統計情報を示す。

ナイジェル・コリアー　Nigel Collier
マンチェスター工科大学卒業後、東芝フェローとして来日。過去12年にわたりテキストマイニングを用いた大容量テキストからのより効率的な情報発見に尽力してきた。世界健康安全保障に関するG7諮問委員会の委員も務める。

川添　愛（かわぞえ・あい）
大学院にて言語学を専攻後、国立情報研究所で自然言語処理やオントロジーの研究に従事。文系の学問と情報科学の接点に興味があり、現在は津田塾大学にて文理融合のための活動を行っている。