3.5.3　ユニバーサルコミュニケーション研究所　情報分析研究室

室長　　鳥澤健太郎　ほか 13 名

ネット情報の意味を深く分析する

【概　要】

インターネット上には膨大な情報が存在し、非常に多くの人々が検索エンジンなどを用いてそうした情報にアクセスしている。一方で、情報の間には様々なつながりがあり、本来はそうしたつながりをきちんと見ていくことで初めて本当の意味で情報を活用することが可能になる。例えば、ある出来事を示す情報があったとして、その出来事の帰結（例えば、ある事件がどのような帰結をもたらすか）や原因（ある事件が起きた原因は何か？）が分かれば、将来の潜在的リスクやチャンス、さらには将来に向けての教訓を知ることにつながり、意思決定をする際に有効である。また、最初に見つけた情報と、帰結、原因といった情報の間の関係はインターネット上で明示的に書かれているとは限らず、情報システムが「考えて」、仮説として原因帰結をユーザに提供する必要もある。例えば、地球温暖化の潜在的な帰結には非常に様々なものがあり、その中には未だ誰も検討していないが、将来には現実的な脅威となり得るようなものもあるであろう。こうした将来リスクを前もって調べる、あるいは予測して、それによるダメージを軽減するためには、少なくとも現状の情報アクセス技術は無力であり、より深い情報の分析や、先に述べたような「仮説」の生成が必要となる。現在、我々は最終的には数十億件の最新の Web 文書などを対象として情報の高度な分析を自動化し、次世代 Web 情報分析システム WISDOM X として一般公開することを目標として研究開発を進めている。

【平成 25 年度の成果】

次世代 Web 情報分析システム WISDOM X（昨年度までの呼称 WISDOM 2013 から改称）は様々な質問に対して、数十億件の Web ページから回答を発見したり、やはり Web ページの情報をもとに仮説を生成したりするシステムである。情報源となる数十億件の Web ページは毎日数千万件のペースで更新され、常に最新の情報をもとに回答や仮説を提示可能である。平成 25 年度はこの公開に向けて、その改良、拡張を行った。以下ではその概要について述べる。

【WISDOM X の分析機能の精度向上及び機能拡張】WISDOM X の中心的な機能の 1 つに「未来分析」機能がある。これは例えば、「地球温暖化が進むとどうなる？」といった質問を与えると、Web ページから、地球温暖化の帰結、例えば、「花粉症患者が増える」「プランクトンが減少する」といったものを抽出し、さらにそれらを組み合わせることで Web ページに記載のない仮説までをユーザに提示する機能である。仮説としては、例えば、「地球温暖化が進むと海水温度が上昇する」「海水温度が上昇すると（大腸菌の一種である）腸炎ビブリオが海中で増加する」「腸炎ビブリオが増えると食中毒が増加する」といった Web 文書から発見された 3 つの因果関係を組み合わせ、「地球温暖化が進むと食中毒が増加する」といった仮説を生成する。この仮説の例は入力となった Web 文書中には記載がなかったが、その後、気候変動に関する著名な科学雑誌でバルト海における進行中の事実として報告された。未来分析の究極の狙いの 1 つは、そうした先端の研究者ですら新規な発見として報告するような事実を仮説として前もってユーザに提示することである。平成 25 年度においては、こうした未来分析機能の高精度化を図った。具体的には仮説生成時に今まで考慮してこなかった文脈情報を考慮する手法の開発と新規な機械学習アルゴリズムの導入による因果関係抽出の高精度化である。この結果、地球温暖化や少子化などの社会問題約 500 個を対象に 5 万件のシナリオを生成した際、出力されたシナリオの 68％について、3 名の被験者中 2 名が各社会問題の非専門家ではあるものの、シナリオが現実になる可能性はあると判断するレベルに達した。この妥当なシナリオのうち少なくとも約 6.5％前後は入力文書に記載のないことが確実なものであり、意外でありながら妥当であるシナリオが生成されることが示された。

また、昨年度までに WISDOM X とは独立なシステムとして開発した、事象の根拠、理由を回答する Why 型質問応答システムを WISDOM X に導入し、日々収集される Web 情報の更新に対応しつつ、例えば「地球温暖化が進むとプランクトンが減るのはなぜか？」といった複雑な質問への回答が可能となった。また、上述した未来分析機能と連携することにより、「地球温暖化が進むとどうなる」といった質問の回答の中から「プランクトンが減る」といった地球温暖化の帰結を選択し、その理由、根拠（例えば、地球温暖化によって海水の温度、比重が変化し、栄養分に富んだ海洋の深層の水が表層のプランクトンに行き渡らなくなる）をワンクリックで

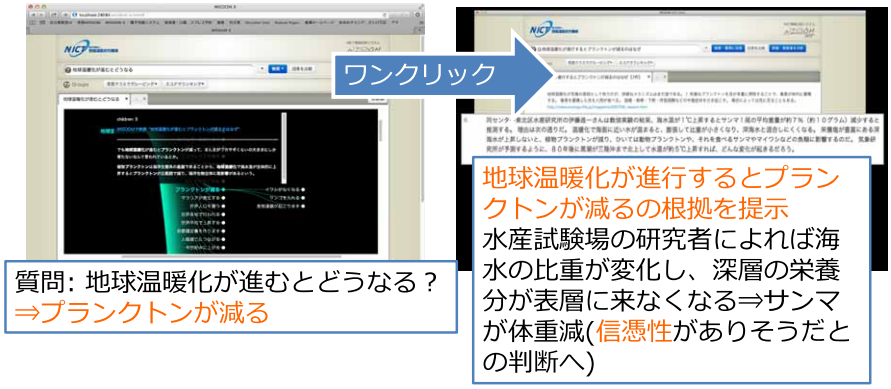

図 1　WISDOM X の未来分析機能と WHY 型質問応答システムの連携

Web ページから発見することが可能になった（図 1）。このプランクトンが減るという例では、Web から当該分野の専門家の記載した情報を発見することができるが、これは、当該事象に対するユーザの理解を深めると同時に、自動生成された仮説も含む未来分析機能の出力の信頼性を判断する材料にもなる。

【WISDOM X の高速化】WISDOM X では分析対象となる Web ページの量が膨大なため、高速な分析処理を行うことがまず必要である。昨年度までは様々な質問に対する回答を検索するためのいわゆるインデックスファイルを大容量ではあるが低速なハードディスク上に格納していた。このような格納方法では、質問によっては回答が出力されるまで数十秒かかるケースがあったため、インデックスファイルをメモリ上に格納する拡張を行った。まずインデックスファイルは巨大なため、多数のサーバーでの並列処理を前提としても、そのままではメモリ上にすべてを格納することは困難となることが予想される。従って、まず質問応答で有用な情報をより詳細に特定し、メモリ上に格納できるサイズまでインデックスファイルを縮小してメモリ上に格納する手法を開発した。なお、こうした変更は日々新たに数千万ページオーダーで収集される Web ページに関する情報の更新が行える形で行われた。この結果、インデックスファイルへのクエリ 1 回あたりの処理時間を数十 ms から数十 μs オーダーへと高速化した。質問応答では、こうしたクエリを質問 1 つあたり数十回から数百回繰り返すため、レスポンスタイムは大幅に改善した。

また、WISDOM X では前述したように、日々、数千万件の Web ページを収集し、多種多様な言語処理プログラムによる解析を行った後、様々な質問に回答を与えるためのインデックスファイルに解析結果を格納している。こうした規模の Web ページの収集、解析を高い効率で安定的に行うことはそれほど容易なことではない。必要な規模を達成するためには、多数のサーバー上で各種の言語処理プログラムを並列実行する必要があるが、こうした並列実行を高い効率で行うには、高度なプログラミング技術が要求され、また不具合の発生の可能性も高い。我々は今年度、低コストでこうした処理を実現するため、基盤的ソフトウェア、つまりミドルウェアである Rapid Service Connector（RaSC）を新規に開発し、WISDOM X に導入した。RaSC は様々な分析プログラムや言語処理プログラムを柔軟かつ高速にネットワーク上で接続し、それらを効率よく並列実行することを可能にする。なお、RaSC はすでにオープンソースとして公開されている（公開サイト: https://alaginrc.nict.go.jp/rasc/）。

【言語資源、つまり知識ベースの拡充】WISDOM X は「地球温暖化が進行するとどうなる？」や「地球温暖化が進行するとプランクトンが減るのはなぜ？」といった単語 1 つで回答できないタイプの質問だけではなく、「セシウムを含むものには何がある？」あるいは「ナノテクノロジーによるビジネスは何ですか？」といった単語 1 つで回答できる質問に対する回答も瞬時に数百個提示する。こうした処理においては、例えば「X が Y を含む」という言語的パターンが「Y の入った X」というパターンとほぼ同義であるという常識的知識が必要となる。こうした同義のパターンは膨大なものがあり、人が手で網羅的に書き表すことは不可能に近い。情報分析研究室では、そうした常識的知識を大量に含んだ知識ベース（言語資源とも呼ぶ）の自動構築の研究を長期にわたって継続してきており、今年度もそうした自動構築手法の精度及び網羅性の改善を図った。特に、今年度は、確かに人が知識を網羅的に書き表してシステムに与えるのは難しいにしても、特別に設計された特殊な辞書を一定量、人手で構築し、知識ベースの自動構築アルゴリズムに事前知識として与えることにより、様々な知識ベースの自動構築の精度、網羅性が劇的に向上することを示した。今後、WISDOM X ではこうした辞書も利用しつつ、知識ベース、分析の改善を図っていく予定である。