2.8 コンテンツ・サービス基盤技術

2.8.1 知識の構造化に関する基盤技術の研究開発

人間の知識の多くは、日本語や英語等のいわゆる自然言語で表される。したがって、Web などに存在する対象の自然言語で書かれた文書から知識を抽出し、整理する手段が存在すれば、知識の構造化が可能になるということになろう。しかしながら、自然言語にはそれ固有の問題があり、そこから知識を抽出すること自体が技術的課題となる。代表的な問題をいくつか挙げると、まず、自然言語においては同様の事象を全く異なる表現で表すことが可能である。例えば、「地球温暖化の原因は二酸化炭素である」という表現と「二酸化炭素は地球温暖化を引き起こす」という2つの表現は「地球温暖化」「二酸化炭素」の2つの名詞句を除けば、まったく異なる表現であるにも関わらずほぼ同義である。自然言語で書かれた知識を抽出、構造化するには、少なくとも、上に挙げた「A の原因は B である」と「B は A を引き起こす」といった全く異なる表現でありながら、ほぼ同一の意味を持ち、また、同一のタイプの知識を表現するものを同一視できなければならない。こうした同義でありながら全く異なる自然言語の表現を認識する技術は、自然言語処理の分野においては「言い換え認識」と呼ばれる重要な未解決課題とされている。

また、語の意味的な分類も重要である。例えば、「イチローのホームラン」と言えば、「イチローが打ったホームラン」の意味であるが、「イチローのユニフォーム」と言えば、「イチローが打ったユニフォーム」を意味する訳ではない。こうした意味解釈の差は名詞「ホームラン」と「ユニフォーム」の違いによっていて、ホームランとユニフォームという単語がどのように意味的に分類されるかに依存していると考えることができる。こうした単語の意味的分類は、シソーラスの構築など、コンピュータが存在する以前から人手で行われてきたが、分類された単語の量、あるいは分類の細かさなど、まだまだ課題が多く、例えば、Web 上に存在する大量の文書から知識を抽出するには不十分であった。

(1) 自然言語処理に基づく知識の構造化

第2期中期計画においては、NICT の言語基盤グループにおいて、こうした自然言語処理における課題の解決に取り組み、平成22年度には、Web 文書等の情報をもとに、構造化された知識として250万語の間の意味的関係を含む概念辞書を自動構築した。また、そうした構造化された知識を利用するアプリケーションとして、スマートフォンに入力された音声での質問にほぼリアルタイムで回答を列挙するシステム「一休」を開発した。一休は6億件の Web ページから質問への回答を抽出する（図2.8.1）。この回答抽出は概念辞書等の知識や概念辞書を自動構築するために開発されたその他の技術を用いて、柔軟に質問に回答するものであり、また、一休自体が6億件の Web ページの知識をオンラインで構造化し利用可能にするものと考えることもできる。

図2.8.1　音声質問応答システム「一休」

処理対象が6億ページの Web 文書と大量であることもあり、非常に意外でありながら、有用な回答を得られる。例えば、「デフレを引き起こすのは何ですか」といった質問の回答としては、意外なことに日本を代表する自動車メーカーの名前が提示された。これはあるブログ中に「＜その企業＞が、巨額の利益を内部留保にまわしたため、総需要が縮小し、デフレを悪化させた」という記述があったのをシステムが発見したからである。この回答を発見するプロセスを展示会等でデモした後、ある著名な経済雑誌でほぼ同主旨の記事が掲載された。つま

り、発端はブログ記事に過ぎなかったわけであるが、経済雑誌で取り上げられるほどの信憑性、インパクトがある回答を先取りで発見したということになる。なお、質問は「デフレを引き起こすもの」を問うていたのに対して、ブログ中の記事では「< その企業 > がデフレを悪化させた」と記述されていた。一休を支える技術の代表的なものが先ほど挙げた「言い換え」の自動認識技術であり、この技術により「引き起こす」と「悪化させる」という字面上全く異なる表現が非常に類似した意味で使われているということが自動的に認識された。こうした技術及びそうした技術によって作成できる辞書は、人の常識、常識的行動をとらえる上で非常に重要である。

また、図2.8.1では中国からの輸入品に依存している企業を複数個の質問を行うことで特定しているが、通常の商用検索エンジンでこうした情報を見つけるには大量の文書を読む必要があり、多大な手間がかかる一方、一休では非常に容易にこうした情報を特定できる可能性を示している。

また、一休には、Web上に書かれた知識に関する推論技術が導入されており、そもそも入力となるWeb文書に（少なくとも直接的に）書かれていないが、妥当である可能性が高い知識を質問の回答として提供することが一部可能になっている。今後こうした推論技術を更に拡張させることによって、Webを単に多様な情報が記載された情報源・データから、いわば「考える主体」へと進化させることが可能となろう。

(2) 成果の公開

また、概念辞書の一部やそれを構築するためのツールやサービスはNICT以外の組織にも提供した。より具体的に述べると、Wikipediaから語（例：情報通信研究機構）とその上位概念を表す語（例：研究所）の間の関係を大量に抽出できる上位下位関係抽出ツールをフリーソフトウェアとして平成22年10月に公開した（http://alaginrc.nict.go.jp/hyponymy/）。この語とその上位概念の間の関係は、上位下位関係と呼ばれ、その語の単語の分類を表すが、同時に非常に基本的な意味的関係と見なすこともできる。その他の意味的関係、例えば、「カビ」と「アレルギー」の間に成立する因果関係等の関係に関しては、6億件のWebページからユーザのリクエストに応じて必要な関係を取得することが可能な「意味的関係抽出サービス」としてALAGIN（詳細はhttp://alagin.jpを参照のこと）を通じて会員に提供している。また、類似する意味を持つ可能性が高い語の対を、それぞれの語の周辺の文脈の類似性によって求めてデータベースにまとめあげた「文脈類似語データベース」や、含意関係を持つ動詞の対を集めた「動詞含意関係データベース」、更には、こうしたデータベースをシステム開発者が自ら開発しているシステムで利用するため、特定の意味を持つ単語の集合を半自動的に収集できる「カスタム単語集合サポートサービス」なども平成21年以降、相次いでALAGIN会員に提供している。これらのツール、サービス、データベースは、大量に存在するWeb等の文書から知識を抽出し構造化するシステムととらえることもでき、例えば、ニフティ株式会社のレシピ検索サービス等で実際に活用された他、それらの拡張が前述した一休に組み込まれている。

以上に述べてきたように、第2期中期計画においては、テキストに書かれた知識の構造化を行うため、大量のWeb文書から知識の構造化を行うための手法や、そのための各種データベースやサービスを開発した。また、音声・言語処理に関係する各種技術、リソース等の共有を念頭に第2期中期計画期間中に設立されたALAGINは、企業や大学の研究者等、会員数は211に達し、NICTで開発されたデータベースやサービスに関してNICTとALAGIN会員とで結ばれた利用許諾契約は平成23年3月時点で、合計445件を数え、社会への還元も順調に進捗した。

2.8.2 情報の信頼度評価などに関する基盤技術の研究開発

平成18年度に開始した「Web情報の信頼性分析」に関する研究は、情報検索エンジンでは困難な情報の信頼性について、ユーザの判断を支援するための情報分析エンジンの構築を行うことを目標に定めた。そのコンセプトを実現するためのWeb情報収集基盤を構築し、情報分析システムWISDOMとして平成21年3月に一般公開した。

以下、開発の概要を(1)～(4)に、また、その学術的成果を(5)に記す。

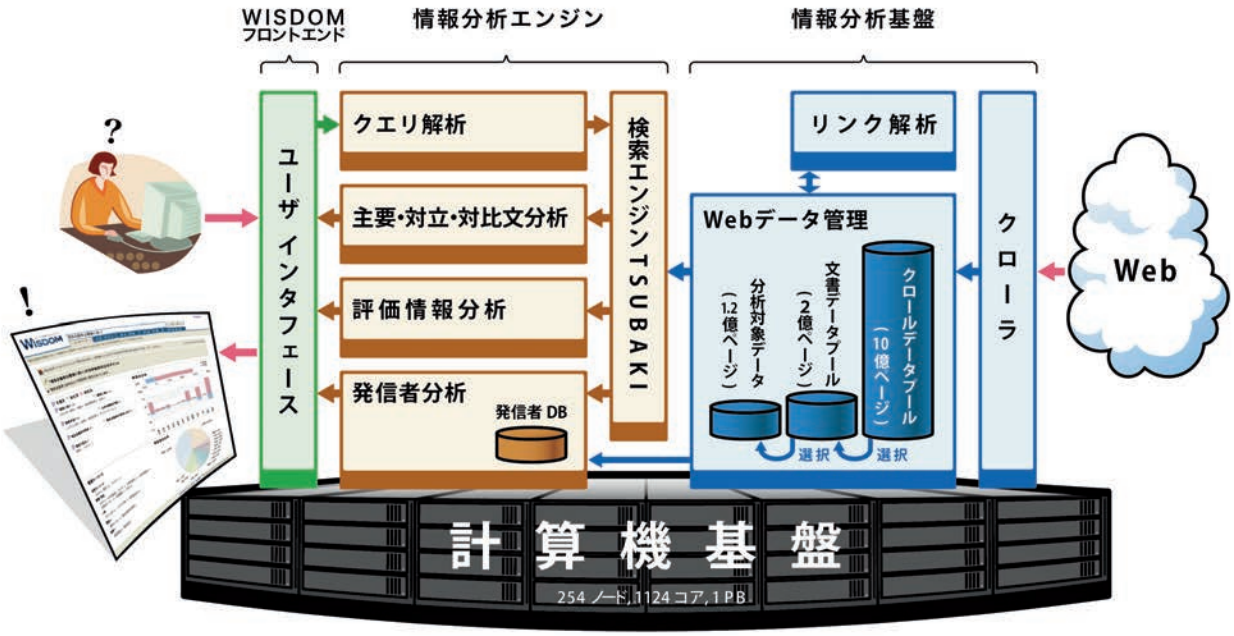

図2.8.2　情報分析システム WISDOM の構成

(1) 情報信頼性分析支援

Web 情報に記述されている情報の信頼性は、ユーザの視点や新たな事実の発見、社会情勢によって大きく変化することがあり、自動的に信頼性の有無を判断することは非常に困難であるため、最終的な判断は人の手によって行われるべきである。そのためには、情報を整理・分類して分析することが必要不可欠であるが、インターネット上で流通する情報は膨大かつ玉石混淆であるため、従来の検索エンジンでは、「誰がどのような意見を述べているのか」ということを整理・分類することすら容易ではなかった。そこで本研究開発では、Web 情報の信頼性分析支援を行うために以下の3つの評価軸を設定した。

a) 情報内容の信頼性
b) 情報発信者の信頼性
c) 情報外観の信頼性

(2) 情報分析システム WISDOM

情報分析システム WISDOM は、以下の3つの機能から構成される。

・情報分析基盤

インターネットから Web ページを収集したデータを管理する Web データ管理機能と、その情報を分析するオフライン分析機能を持つ。

・情報分析エンジン

ユーザの分析要求に基づいて分析する複数機能を持つ。

・WISDOM フロントエンド

分析結果を入力し、その処理結果を整理・分類して表示するユーザインタフェース機能を持つ。

次項以降にそれらについて記述する（図2.8.2）。

(3) 情報分析基盤の構築

情報分析システムの構築においては、分析対象とする Web ページの大規模収集が必要不可欠である。WISDOM の開発と並行して Web 情報を収集するためのクローラの開発や Web アーカイブなどの構築を行った。これらの情報分析基盤は平成22年度には約500万ページ／日の収集を行いつつ、7億ページ規模の Web アーカイブを構築した。

収集された Web 情報は、アーカイブに収納された後、ユーザのクエリ（問いかけ）が与えられなくても分析可能なリンク解析、外観分析、発信者分析などの処理（以下、オフライン処理）が行われ、その分析結果は情報分析基盤の主要な情報として格納される。

a) リンク解析技術

収集した Web ページには Web スパムと呼ばれる無意味なページが含まれている。この Web スパムはコン

テンツスパム、リンクファーム、なりすましといった3種類に大別される。このリンクファームを検出するためにWeb間のリンクを大規模なグラフデータとして抽出し、高密度なサブグラフを抽出した上でスパム判定を行う技術を開発し、実装している。

b) 外観分析技術

Webページは構造化文書であるため、そのHTMLタグの知識を持つユーザが発信する情報は、構造化されて記述されている。一方、スパムページなどにおいては、その構造が整理されていない場合も多く見受けられ、外観的な特徴として不整合性がある場合が多い。このようなWebページの外観情報もWebページの信頼性と関係しており、WISDOMでは文書構造の解析の結果、あるべき情報が欠落していないか等の外観情報を分析する。

c) 発信者分析

情報の信頼性判断には、「誰が発信しているのか」等の発信者情報が極めて重要な情報となる。専門家が発信している情報と、専門家以外が発信している情報では、多くの場合専門家が発信している情報の信頼性が高いと考えられる。WISDOMにおいては、Webページの情報の内容及び、その公開について責任を有する人物や団体などを含む実態を発信者と定義して、サイト運営者と著者に分類する。さらに、情報発信者クラスとして6種類に分類し、各Webページの発信者情報の分析結果を整理する。これらの分析においては、情報発信者と想定できる発信者候補が記述されている文（以下、候補とする）を抽出するために以下を手がかりに分類する。

① 情報源全体における候補出現頻度
② 候補が出現するページの頻度
③ 候補が出現する文書の種類
④ 構成語の品詞属性
⑤ 先頭形態素・末尾形態素
⑥ 形態素数
⑦ ページ内位置
⑧ 著作権表示由来か否か

d) 情報分析エンジン

ユーザが入力した分析要求クエリに応じて、情報分析基盤に格納されたWeb情報に対して分析を行う情報分析エンジンは、クエリ解析の後に主要・対立・対比分析や評価情報分析を行い、オフライン処理結果の情報と合わせて出力する。

e) クエリ解析

WISDOMのクエリは自然文が与えられることを想定しており、その解析によって何に対して分析を行うのかを決定する。そのため、入力されたクエリはトピックとサブトピックに分類され、評価表現分析には、そのトピックとサブトピックが渡され、主要・対立・対比表現には、トピックが渡される。

f) 主要・対立・対比分析

クエリ解析によって抽出されたトピックに関する関連キーワード及び、主要・対立・対比文を対象となるWebページ集合から抽出する。そして、述語1つとそれにかかる1つ以上の自立語列を抽出した述語項構造を抽出した後に、同義の述語項構造や包含関係を分析して意味的に同一のものを集約し、それぞれ主要・対立・対比の分類を行う。

g) 評価情報分析

クエリ解析によって得られたトピックに関する肯定的・否定的な意見や評価をWeb文書から自動的に抽出・分類する。WISDOMでは、「感情」「批評」「メリット」「採否」「出来事」「当為」「要望」の7種類の分析をしている。収集したWeb情報から1トピックあたり200文について評価情報を人が評価した上で、タグ情報として付与し、2,000文の評価情報タグ付きコーパスを作成し、機械学習用の教師データとしている。このコーパスを元にして、サポートベクトルマシン（SVM）による分類を行い、関連度の高いものを評価情報として出力している。

(4) WISDOM フロントエンド

WISDOMは、従来の情報検索エンジンと同様の手軽さで、分析要求を自然文で入力できる。例えば、図2.8.3のように「電動歯ブラシは歯に良い」という文を入力すると、以下の様に分類されて表示される。

① トピックの定義
② 分析結果の注目ポイント
③ 主要・対立・対比文
④ 発信者毎の意見の分布
⑤ 関連キーワード
⑥ 発信者の分布
⑦ 主な発信者と主な意見

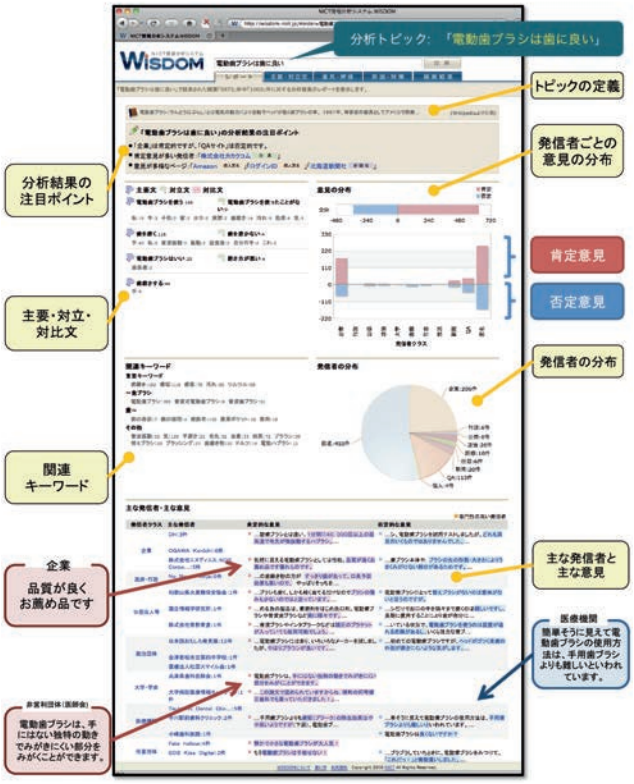

図2.8.3　WISDOM の利用画面例

(5) 学術的成果

情報の信頼性分析を目標とした情報分析エンジンの開発は極めて挑戦的な研究開発課題であり、実用レベルにまで高めた技術として一般公開したことは社会にも高く評価された。学術的な新規性のみならず、社会に貢献したことが評価され、平成23年3月に第56回前島賞、同年5月に第43回市村学術賞 貢献賞を受賞している。さらに、この WISDOM の研究開発は NICT における情報分析技術という研究分野として確立された。平成23年度以降の研究成果については、**2.8.4**に記載する。

2.8.3　ナレッジクラスタ形成技術の研究開発

ナレッジクラスタシステムは、従来のインターネットをより知的な情報獲得と分析の環境へと進化させることを目的に、様々な分野や組織で個別に蓄積されてきた情報を横断的に連結・統合することで世界規模の知識ネットワークを構築するための基盤技術である。ナレッジクラスタシステムは、グリッド基盤と異分野相関検索エンジンによって構成される（図2.8.4）。

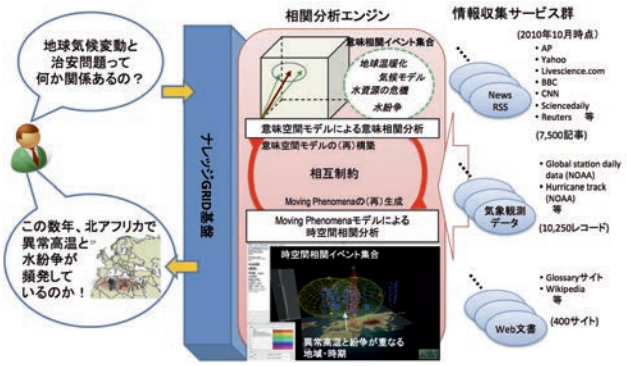

図2.8.4　ナレッジクラスタシステムによる相関検索

(1) 相関エンジンの開発

従来の Web のようにデータを配信・共有するだけでなく、特定のデータを意図的に集めたり、集めたデータの中から様々な話題や出来事に関する情報を抽出したり、抽出した情報を様々につなぎ合わせたり、つなぎ合わせられた情報を検索したり閲覧したりする機能を備えている。例えば、地球気候変動と治安問題のつながり（相関関係：correlation）を知りたい場合、従来の検索エンジンでは“地球気候変動　治安問題”というキーワードを含む Web ページを検索し、検索結果の内容をユーザが直接自分で判断するしかなかった。一方、我々が開発したシステムは、同じ質問に対し、例えば「この数年、北アフリカで異常高温と水紛争が頻発している」というような、キーワードで直接指定されていない内容にまで範囲を広げて相関のある情報を見つけ出すことができる。相関検索エンジンを使って、Web ページはもちろん、ニュース配信や気象観測データまで多岐にわたる大量のデータを対象に、それらの意味的なつながりと時空間的なつながりのあるデータを発見する。その際、従来のように、予め共通の辞書を用意しておく必要がないため、異種・異分野のデータの相関も拡張性高く発見することができる。相関検索エンジンは、時間、空間、主題に関する様々な種類の特徴量を使ってデータを索引付け、問い合わせ処理の際は、相関を発見するのに最適な特徴量の組み合わせ（相関の文脈と呼ぶ）と、それらを用いて高い相関を示すデータ集合を同時に検索する。その結果、先程の例のように、「この数年（時間の特徴量）、北アフリカで（空間の特徴量）異常高温と水紛争が（主題の特徴量）頻発している」という文脈の中で高い相関を示すデータ集合を発見することができる。NICT は、意味空間モデルと、Moving Phenomena 時空間モデルによ

る相関分析手法を平成20年度に提案したが、いずれも多次元空間における部分空間選択により最適な特徴量の組み合わせと相関データ集合を柔軟性高くかつ効率的に見つけ出すことができることが特徴である。

(2) ナレッジクラスタシステムの構築

ナレッジクラスタシステムの全ての機能は、グリッド基盤の上に実装されている。アジアやヨーロッパなど世界中の拠点をつないだ広域グリッドネットワークを構築し、その上にサービス指向アーキテクチャに基づいたグリッド環境を平成20年度に実現した(図2.8.5)。

グリッド基盤上には、データ収集と情報抽出、相関分析、ユーザインタフェース・インタラクションに関する多種多様なソフトウェアサービスが、各拠点で並行して開発され配備されている。これらのサービスを組み合わせ連携させることで、様々な次世代Webアプリケーションを実現することができる。平成21年から22年にかけて、Webをハイパーリンクとは無関係に相関関係に沿って閲覧するLink-Free Web Browsing、場所や時期、テーマに応じて相関性のある情報を次々と推薦する観光案内システム、企業や組織の製品・サービス・技術に関する情報をWebサイトや特許から幅広く収集し、様々なテーマの下でマッチメイキングを行うパートナーシップマネジメントシステムなど、様々なアプリケーションを開発してきた。グリッド基盤には、アプリケーション開発のためのサービス連携機構と開発言語を実装した。また、アプリケーションでの使われ方に基づいて役立つサービスを検索する機構(サービス発見エンジン)を開発した。

2.8.4 情報分析技術の研究開発

いわゆるインターネット上の情報爆発には収束の気配はなく、インターネット上に存在するいわゆるBig Dataから価値を創出することは全世界的に課題と見なされている。こうした観点から、NICTにおいてもインターネット上の大量の情報を分析する技術、方法論の研究が第2期中期計画に開始され(**2.8.2**)、第3期中期計画においても、インターネット上のテキストを深く意味的に分析し、それらの価値ある組み合わせや分類を発見するWISDOM Xという大規模なWeb情報分析システムの開発を行っている。このシステムは日々1千万件以上の速度で収集される全体では数十億件単位のWebページを対象として、様々な言語処理、情報分析サービスを連携させ、従来にない高度な情報の分析を可能にするシステムであり、数百台のサーバーから成るクラスタ上で稼働していて、現在、一般公開に向けて改良を進めている。

また、平成23年の東日本大震災後、被災状況の把握が非常に困難であったという反省に基づき、Twitter上の災害関連情報をリアルタイムで分析し、自治体、NPOなどの救援サイド、更には被災者に被災状況を分かりやすく提供する対災害情報分析システムの開発も、耐災害ICT研究センターと共同で実施している。この

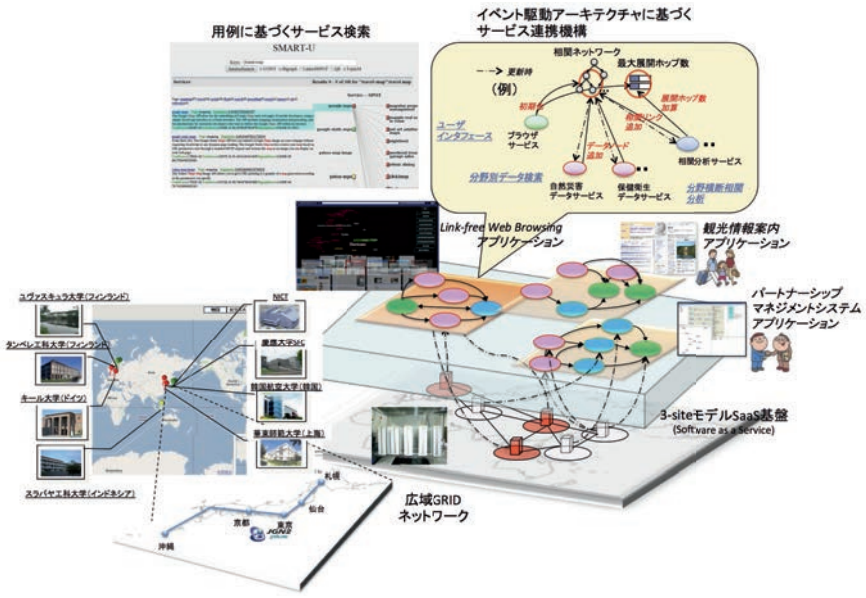

図2.8.5 サービス指向アーキテクチャに基づいたグリッド環境

システムは平成26年11月に公開し、将来起きる可能性のある大規模災害における救援活動で活用していただけるものと考えている。以下では、これらの2つのシステムの概要について述べる。

(1) WISDOM X の研究開発

WISDOM X は、基本的に、自然言語での質問に対して Web 上の情報をもとに回答を提示するシステムである。例えば、「少子化が進行するとどうなる？」といった質問を入力すると、Web 上に記載された情報に基づいて、少子化の帰結となる可能性のあるフレーズを多数提示する（図2.8.6）。WISDOM X の特徴は、商用検索エンジンのように、入力された表現を含む Web ページを少数提示するのではなく、回答となる可能性のある表現をピンポイントで提示すること、また、そうした回答の候補を多数（場合によっては数百件から数千件）瞬時に提示することである。WISDOM X の開発の狙いは、価値のある未知の情報をユーザに発見してもらうことであり、そのためには、多数の回答を概観できることが望ましい。例えば、前述の「少子化が進行するとどうなる？」という質問に対しては、「不況になる」「労働者人口が減る」「過疎化が進む」といった常識的な回答から「少人数教育が増加する」「ペットブームになる」「国公立大学志向が強まる」といった普段、新聞やマスコミ等を見ているだけではなかなか想定しづらい回答までが多数提示される。例えば、少子化の対策を検討したり、少子化時代のビジネスチャンスを検討しているユーザにとってこうした回答の一覧は他では入手できないものであり、価値ある情報である。

また、これらの回答をクリックすると、更に帰結を深堀りして調べることが可能になる。例えば、「地球温暖化が進むとどうなる？」という質問の回答の中に「海水温が上がる」というものがあり、それをクリックすると、海水温が上がった結果「腸炎ビブリオ（大腸菌の一種）が（海中で）増加する」という更なる帰結が提示される。さらにそれをクリックすると、今度は、腸炎ビブリオが増えると、「（シーフードによる）食中毒」が増えるという結果が提示される（図2.8.7）。ちなみに、海水温が上昇した結果、腸炎ビブリオが増加し、食中毒が増えるという一連の出来事は、地球温暖化と無関係に毎年夏になると生じる現象であるが、地球温暖化から、腸炎ビブリオ、食中毒の増加に至る一連のシナリオは WISDOM X が Web 中の情報を組み合せて自動的に生成した仮説であり、入力された文書中にこのシナリオは記載されていなかった。一方で、このシナリオは2007年に収集した一連の Web ページから生成されたものであるが、その後、著名な学術誌でバルト海において現在進行中の事実であることが報告されている（Baker-Austin et al., Nature Climate Change 3, 73-77（2013））。大量の Web ページをもとに、こうした素人では全く想像もつかず自明でない仮説は、様々な意思決定の際に将来起こり得る事象を広く見渡してリスク、チャンス等を検討し、バランスのとれた決定を行う上で価値ある情報であり、システムは今後非常に有用なものになると考えている。

前述の「～するとどうなる」という形の質問に回答する機能を我々は未来分析機能と呼んでいるが、

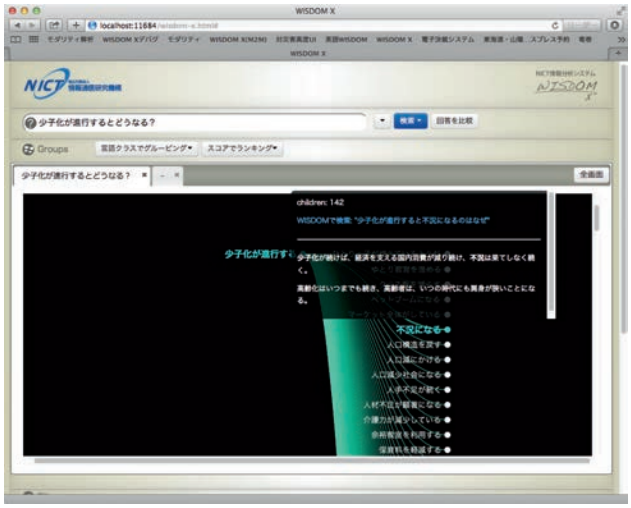

図2.8.6　質問「少子化が進行するとどうなる？」に対する WISDOM X の回答

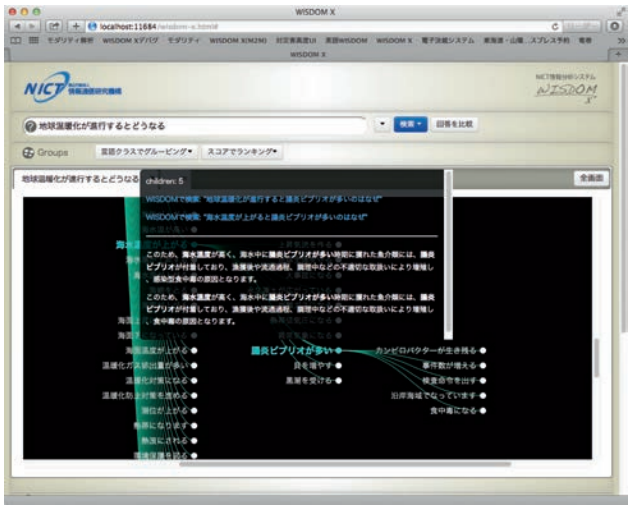

図2.8.7　地球温暖化にまつわる仮説的シナリオ

WISDOM X にはこれ以外にも「ナノテクノロジーによるビジネスはなにか？」といった名詞1つで回答できる質問に対するファクトイド質問応答機能、「なぜ日本の農業は弱いか」といった理由を問う質問に文章で回答する WHY 型質問応答機能（図2.8.8）などが備わっている。これらの機能は連携することも可能であり、例えば、未来分析機能で「地球温暖化が進むとプランクトンが減る」といった情報が得られた際に、その理由・根拠をワンクリックで WHY 型質問応答機能により得ると言ったことも可能になっている。WISDOM X ではこのような異なる質問応答機能の組み合わせで、物事を考える上でのヒントをより広く提供できるようになっており、平成27年度中に一般公開を行う予定であり、広く利用されることを期待している。

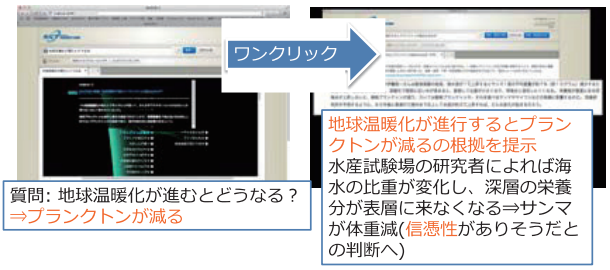

図2.8.8　WHY 型質問応答による仮説の検証

(2) 対災害情報分析システム DISAANA

冒頭で述べたように、WISDOM X の技術を一部転用する形で、災害時の Twitter 情報をリアルタイムで分析し、その結果を自治体、NPO、被災者等に分かりやすく提供する対災害情報分析システムの開発も行っている。例えば、「宮城県で不足しているものは何か？」「宮城県で透析が受けられるのはどこか？」といった質問に対して回答を提示する（図2.8.9）。また、回答結果を地図上に表示したり、災害時を意識した意味的分類を用いて分かりやすく提示できるインターフェースが実現されている。このシステムで得られた回答を通常の商用検索エンジンで短時間に発見することは困難であり、緊急時には必要不可欠なシステムになると考えている。

また、Twitter 等の SNS では災害時にデマ・流言が問題となるが、これに対処するために、回答と矛盾する情報が得られた場合にはそれも提示し、ユーザが情報の信憑性を判断する際の材料を提供することもできる。これは国立大学法人東北大学と共同で開発している「言論

図2.8.9　対災害情報分析システム

マップ」と呼ばれる機能であり、例えば、「放射能に効くのは何か？」という質問に対して「イソジン」という回答が発見された場合、「イソジンが放射能に効くというのはデマです」といった、回答と矛盾する情報を回答と共に提供するものである。こうしたいわゆるデマの訂正は東日本大震災時にも実は多数発信されたが、それを効率的に検索する手段がなかったため、有効活用はされなかった。また、こうした機能が被災時に機能することで、より積極的にデマの訂正が行われることも期待できる。このシステムは平成26年11月に一般公開を行っており、広く利用されることを期待している。

以上、第3期中期計画で開発している情報分析技術の概要について述べてきた。これらの技術はいずれも10年前までは空想の域を出なかったような技術であるが、Web 上で大量のテキストが入手可能となり、機械学習のような先進的な技術を利用することで実現可能となった技術である。こうした膨大なテキストの利用方法の可能性はまだ十分に汲み尽くされてはおらず、今後も様々な先進的サービスを研究開発する予定である。

2.8.5　情報利活用基盤技術の研究開発

近年、様々な組織や機関、個人が収集したデータを公共財として自由に利活用できるようにするオープンデータの取組が世界的に広まってきている。オープンデータの普及に伴い、多種多様なデータを利用できる環境が整いつつあり、これらを横断的に活用しコネクションメリッ

トを発揮する技術へのニーズが高まってきている。

(1) 分野横断相関検索技術の研究開発

平成24年度から研究開発を行っている分野横断相関検索技術 Cross-DB Search は、オープンデータの中でも特に実世界の状況を反映した様々なセンサーデータや科学データを対象に、地理的、時間的、概念的、及び引用参照の4つの相関性からデータ間のつながりを複合的に評価し、クエリにヒットするデータに対し相関の高いデータ集合を見つけ出す。例えば、大気品質の1つの指標である‘PM2.5’をクエリとして与えると、PM2.5の観測データだけではなく、それらの周辺の地域や時期に作成されたデータや、PM2.5と関連性のある分野のデータ、更にはそれらと一緒に引用参照されることが多いデータを芋づる式に検索し、これらの中から相互に相関の高いデータ集合を発見する(図2.8.10)。

図2.8.10　異分野相関検索 Cross-DB Search

Cross-DB Search では、平成23年度から平成26年度までに、ICSU(国際科学会議) World Data System に登録されている40分野100万データセットを超える科学データや、独自に収集した50種類以上のセンサーデータを対象に、相関検索に必要なメタデータを抽出し索引付けしている。この Cross-DB Search を用いて、様々な調査研究のために幅広い分野のオープンデータを収集、整理、保存、公開を行うデータキュレーションへの応用に取り組んでいる。そこでは、相関性を検証するためのデータセットの作成と、データセットを分析し様々な観点から相関性を発見することを繰り返し行いながら、相関分析と検証データを同時に絞り込んでいく。こうしたサイクルをいかに迅速かつ効率的に回すかが、特にデータを中心とする科学では重要とされ、Cross-DB Search はその効率化に役立つ。

(2) 大規模情報可視化技術の研究開発

様々な分野のデータの横断的な関係性を把握する上で、情報可視化は有効な手段である。NICT では、物理センシングや社会センシング(SNS など)によって得られる様々な種類のセンサーデータや、幅広い分野の観測データを蓄積した科学データアーカイブなど、実世界を反映したオープンデータを中心に、実世界の様々な出来事(イベント)の分野横断的な相関性を視覚的に分析する情報可視化技術の研究開発を行っている。平成24年度から研究開発を行っている STICKER は、地理空間と時間軸から構成される3次元空間上に様々なセンサーデータを表示し、様々な可視化手法を用いてそれらの動きや変化の相関性を視覚的に分析する。STICKER の特徴は、単にデータを可視化するだけでなく、様々なデータの中から相関の高いデータの組み合わせを発見し、かつ高い相関を示すデータ部分を抽出することである。図2.8.11の例では、まず PM2.5のデータの分布を可視化し、異常に変化している箇所などユーザが関心を持つ部分を指定し、それらと相関の高い他のデータを検索する。さらに検索結果を時空間上に表示する際に、可視化するデータの種類や値域を操作しながら相関が高くなるように調整する。こうしたデータの表示や検索、操作を繰り返しながら相関の高いデータの組み合わせを発見し、後に続く詳細な解析に必要なデータを準備する。これらの作業は、膨大な種類のオープンデータから役立つデータを効率的に絞り込む上で重要である。STICKER では、様々な分野のデータに対し横断的な操作や可視化を可能にすべく、STT(Space、Time、Theme)スキーマでデータを統一的に構造化し、各種データから STT スキーマに変換して、様々な形状の可視化オブジェクトの生成とそれらの視覚的操作、STT 属性に基づくデータの選択や集約、異常値検出などの各種データ操作、及び相関検索を実装している。

(3) 知識・言語グリッドの研究開発

オープンデータの普及は、これまで主に政府や研究機

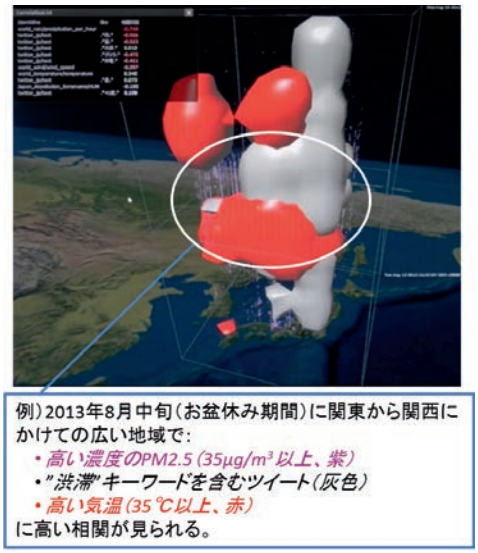

図2.8.11　異分野データの時空間相関可視化分析システム STICKER

関の主導により既存のデータを公開する活動が中心であった。それら初期の取組の成功を受け、今日ではユーザが自らデータを収集、公開し共有する参加型のオープンデータ活動が広まりつつある。このような参加型のオープンデータ活動を支える ICT プラットフォームとして、我々は K-L Grid（知識・言語グリッド）の開発を行っている。K-L Grid は、新世代通信網テストベッド JGN-X 上に構築された広域分散グリッドネットワークであり、現在国内5拠点（東京、北陸、京都、岡山、福岡）のグリッドノードによって構成されている。この上で、オープンデータの収集や共有を協調して行い、情報資産のリポジトリを構築している（表2.8.1）。

また、グリッドネットワークの特徴を生かし、ユーザ独自のノードを参加させることで、ユーザが K-L Grid 上で自身のデータを参加ノード経由で公開したり、他のデータと組み合わせて利用することを可能にしている。さらに、情報サービスの要求に応じてネットワークの構成を自動的に設定する SCN 技術の研究開発も行い、データの収集、加工（フィルタリングや分割・統合など）、スループット等の要件に基づいてネットワークの発見やパス生成、QoS 制御を動的に行う仕組みを実現することで、従来に比べ、より多くの情報サービスをより安定的に稼働させることを可能にしている。こうした取組を普及させるべく、米国標準技術院（NIST）と協力し標準化活動にも取り組んでいる。

表2.8.1　K-L Grid 上の情報資産

分類	種類
物理センサーデータ	地震情報、地滑り危険地域情報、感染症情報、花粉情報、環境放射線水準情報、降雨量情報、積雪量情報、気温情報、風速風向情報、台風情報、犯罪情報、河川水位・雨量情報、潮位情報、インフルエンザ情報、世界災害情報など49種類
社会センサーデータ	Twitterアーカイブ、ジオタグ付Twitterデータ、トレンドキーワードで集約したTwitterデータ、RSSニュース、Googleニュース
WISDOMデータ（含Webアーカイブ）	全文データ、言語解析済みデータファイル、発信者データ、評判データ、係り受けデータ
科学データ	World Data Systemメタデータ（Pangaea, ICPSR, DRYAD, ESDS, ADAなど40分野）
公共データ	Data.govなどの電子政府オープンデータのメタデータ
地理データ	ランドマーク、避難所データ
言語データ	EDR概念辞書、日本語WordNet、WordNet
情報分析ツール	意見評価、一休サービス、文・フレーズ間の意味的関係DB等
翻訳ツール	VoiceTraテキスト翻訳、JServer
テキスト解析ツール	固有名詞抽出、形態素解析器、係り受け解析器
地理情報ツール	Google Geocoding、Yahoo Contents Geocoder、ランドマーク名抽出、郵便番号検索、GeoNLP
音声ツール	VoiceTra音声サービス（認識、合成）、Rospeexクラウド型音声コミュニケーションサービス

75 種類・125 万データセット規模のデータやツール

（2014 年 1 月時点）