検索アルゴリズムを含めた知識情報技術の動向調査

平成 19 年 2 月 28 日

情報・システム研究機構

目次

１．序論

生物情報を扱うデータベースは、いろいろな分野の異なる観点から作成され、また、それぞれ異なる形式で記述されている。また生物情報データベースに含まれるデータ量は計測技術の発展に伴い膨大な量となってきている。また、High Wire Press や PubMed Central をはじめとした文献の電子化・オープン化が進むことで、生物情報として利用可能なテキストや図表の量も増えつつある。さらに、これら生物情報の利用方法自体、ユーザによってさまざまである。このような背景をもつ生物情報のデータベースを統合するためには、情報科学の高度な技術の利用が不可欠である。

本報告書では次世代の生物情報データベース統合に必要となる、あるいは利用する可能性が高い情報科学技術について、聞き取り調査や文献調査によって動向を調べたものである。まず、第２節でデータベースのインタフェイスの要ともいえる検索システムについて述べる。次に、第３節では、膨大なデータにどのように対処していくかという問題に対する答えとなりうる、データマイニングの動向について述べる。さらに、第４節では、データベースの一つの形ともいえる Web2.0 のコンセプトについて紹介する。そして、データベース統合のアーキテクチャとして期待されるグリッドコンピューティングについて第５節で述べる。第６節では、第２節から第４節まで述べた技術を生物情報のデータベース統合にどのように生かせる可能性があるかについて述べる。最後に、この報告を書くにあたって聞き取り調査で得た資料を添付する。

２．検索システムについて

データベースから情報を効率よく取り出すためには、良質の検索システムが不可欠である。これまでデータベースの検索システムは、項目名からの検索あるいはキーワードによる検索がほとんどであった。しかしながら、これらの検索システムを使いこなすには、ユーザがそのシステムについてある程度以上の知識を持ち、さらに、適切に項目名、あるいはキーワードを選ぶスキルを持つ必要がある。

一方、生物情報における検索は、上記のような単純な項目検索やキーワードサーチでは情報の膨大さ多様さゆえに対応が難しく、検索エンジン自体に高度な解析機能、可視化技術が必要となってきている。また、検索対象がローカルなデータベースからグローバルなWWW上のデータへと広がったため、構造化されたデータベース内のデータのみならず、WWW上の非構造化データをも扱える必要が出てきた。

この節では高度な可視化インタフェイスをもつ検索エンジンの例として、連想検索エンジン DualNavi について述べる。また、構造化データと非構造化データをともに扱う技術として、UIMA について述べる。

２．１　連想検索

連想検索とはユーザの意図に基づく連想からのフィードバックを用いて膨大な情報の中からインタラクティブに情報を絞り込む手法である。この手法を実装した検索エンジンがDualNavi である（http://bace.hgc.jp/dn_intro/img/dn_intro0_J.gif 参照)。その特徴は、ユーザの意図に近い文書群から類似文書を検索する類似文書検索エンジンと、その文書群を特徴付ける特徴語グラフの相互連携にあるといってよい。この二つの仕組みの組み合わせにより、文書群を効率的に絞込み、また、絞り込んだ文書の特徴語を選んでさらに文書群を更新して目的の文書群へたどり着くことを狙っている。

さらに、国立情報学研究所の研究グループは、この連想検索の考え方を発展させ、さまざまなデータに応用している。たとえば、文化遺産オンライン(連想検索によって，文化遺産データベースから選んだ品に近いものを提示)、新書マップ、ジンボウナビ、WebCatPlus などのサービスを始めている。さらに、複数のソースでの連想検索を一覧表示可能にしたサービス、想「IMAGINE」を Ajax ベースで開発している（http://imagine.bookmap.info/imagine 参照)。想の考え方は、多種多様なデータソースのデータ群に対して、ユーザが意図する方向へフィードバックをかけていくことで、目的のデータ群へたどり着けるというものである。これらのサービスには連想検索に特化して開発されたエンジン GETA が使われている。GETA はオープンソースで配布されている（http://geta.ex.nii.ac.jp/)。

２．２　非構造化データ管理技術 UIMA

UIMA(the Unstructured Information Management Architecture)とは、IBM 社が提案する、テキスト解析の異なるツールやアプリケーションを統合すると期待される新しい標準化のフレームワークである。このフレームワークに準拠することで、様々な種類の非構造化データに対する検索エンジンがひとつのインタフェイスで扱えることになる。UIMA の概念図を図２．２．１に示す。UIMA は、テキスト検索エンジンなどの TAE(Text Analysis Engine)のインタフェイスの共通フォーマット、および、異なる TAE のインデックスをつなぐための仕組みである CAS(Common Analysis System)からなる。TAE は UIMA のフォーマットに準拠さえしていれば、対象となるデータの種類は問われることなく、UIMA のプラグインとして利用可能である。このことから、UIMA はサーチエンジンの統合を通して柔軟に異なるデータの統合を果たすことが期待されている。実際、米国防衛高等研究計画局(DARPA : Defense Advanced Research Projects Agency)の大規模プロジェクト GALE では多言語文書・翻訳・音声処理を統合的に行うためのフレームワークとして UIMA が採用されている。UIMA は公開されており、また、UIMA 準拠の TAE を作成するためのツールも無償で利用可能である。

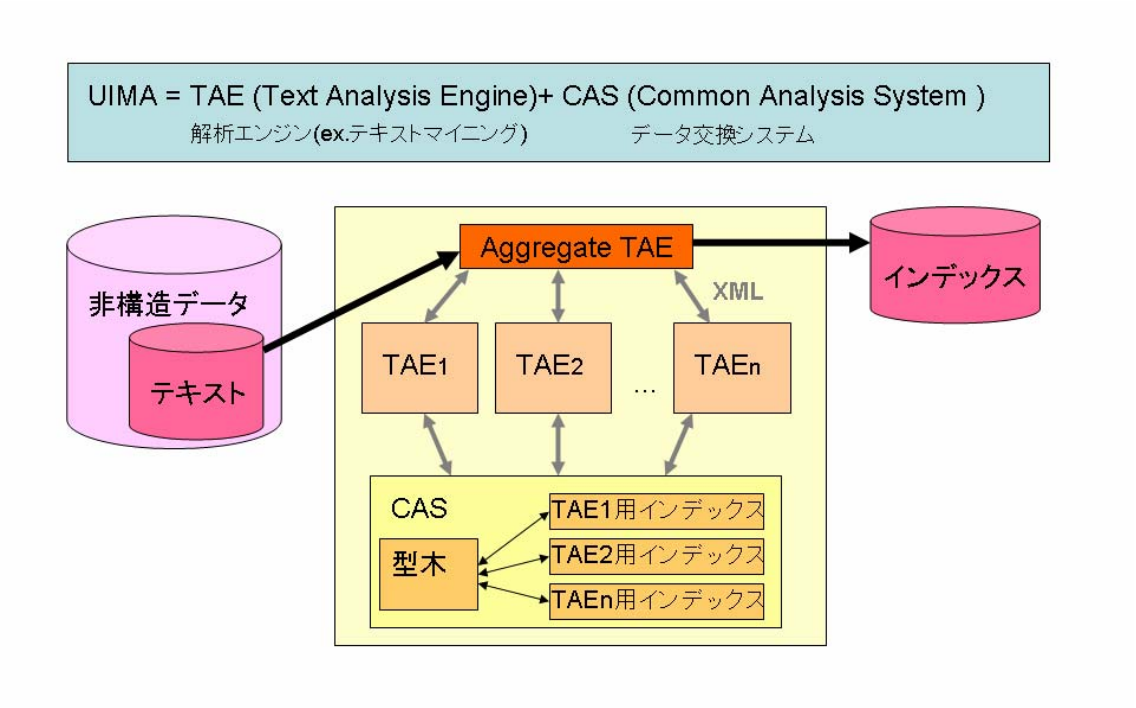

図２．２．１ UIMA(Unstructured Information management architecture)

３．データマイニングについて

３．１　知識発見とデータマイニング

データマイニングとは、大規模なデータやデータベースから隠れた関係性や知識などの情報を帰納的に抽出する技術一般を指す言葉である。マイニングとは「鉱山発掘」の意味であり、蓄積された膨大なデータの山から未知の情報を「発掘する」＝「マイニングする」イメージから名付けられた。データマイニングは多種多様な技術分野を包含している。相関分析や統計的推定などの統計学の技術、ルール抽出や自動分類などの人工知能・機械学習の技術、さらには心理学や認知科学の技術すら含まれることがある。

データマイニングの手法は大規模データからの知識発見のプロセスの一部としてしばしば用いられる。図３．１．１にデータマイニングを用いた、典型的な知識発見プロセスを示す。まず、生データからデータマイニングの対象としたいデータを選択する。次に、ノイズやはずれ値などの除去を行う。データマイニング手法によっては、ノイズに敏感なものもあれば、ノイズに対して影響を受けにくい手法もあるため、このプロセスは使いたいデータマイニング手法についてユーザがある程度理解しておく必要がある。さらに、ノイズなどを除いた洗浄済みのデータに対し、必要な属性を選び、あるいは、使いたいデータマイニング手法の入力形式に合わせて変換を行う手続きが必要である。このプロセスは適

切な属性選択ができるかどうかで、仮説の良し悪し、計算の効率など大きく変化する可能性があるため、データマイニング手法に対する理解と同時にデータそのものについての理解も必要となる。この後、属性選択・データ変換後のデータについて、データマイニング手法を適用してその結果を得る。ここで注意すべきなのは、データマイニング手法で直接得られるのは仮説であり、知識そのものではない点である。したがって、データマイニングで得られた仮説についてユーザが評価し、あるいは検証する必要がある。ここで、ユーザが仮説に満足するか検証できれば、その仮説は知識となり、知識発見のプロセスは終了する。一方、ユーザが思うような仮説が得られない場合は、知識発見のプロセスのどれかに問題がある可能性がある。ユーザは各プロセスにおけるパラメータを変えたり、手法自体を変更したりすることで、新たな仮説を得ることができる。そして、知識発見プロセスは、十分よい仮説が得られるまでこの生データから仮説獲得までのプロセスを繰り返すことになる。

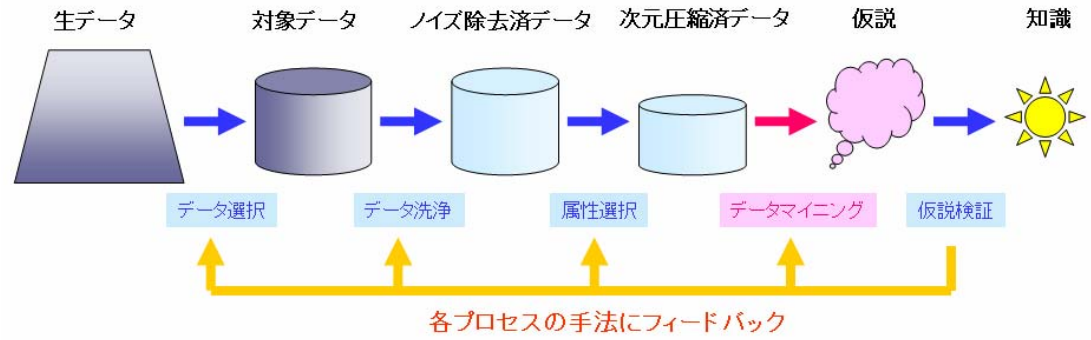

図３．１．１　知識発見プロセスとデータマイニング

３．２　グラフマイニング

データマイニング手法は出力される情報の方向性と入力されるデータの種類から、おおまかに第一世代と第二世代のものに分けることができる。第一世代のデータマイニング手法は、帰納推論によるルール発見、k-means などのクラスタリング、SVM や決定木などの自動分類など、古典的なデータマイニング手法が含まれる。一方、第二世代のデータマイニング手法として、ベイジアンネットワーク、隠れマルコフモデルなどの確率モデルや、グラフマイニングなどの構造データからのマイニング手法、さらに、テキストマイニングやストリームマイニングなどの新しいタイプのマイニング手法が含まれる。本節では、この

第二世代のデータマイニング手法のうち、グラフマイニングを取り上げて解説する。さらに、グラフマイニングの応用例を紹介する。

グラフマイニングは、グラフ構造で表現された大規模データから埋もれた特徴的なパターンを取り出す、頻出パターン抽出問題を解くものが主流を占めている。しかしながら、一般のグラフについては、たった二つのグラフの間の共通パターンですら効率的に取り出せないことがわかっており(最大共通部分グラフ問題が NP 困難であることより)、現実的に頻出パターンを大規模グラフデータから取り出そうとするならば、入力あるいは出力となるグラフの形状に制限をかける必要がある。

近年マイニング対象として注目されているグラフの形状にラベル付木がある。木とは文字通り樹状の形をしたグラフで、厳密にはサイクルを持たないグラフを指す。ラベル付木とは木の頂点にラベルとよばれる離散値がついたものをいう。木の頂点のうち一つを根として、根から他の頂点へ方向性をつけることによってできる有向グラフを有向木という。さらに、有向木の辺の間に順序が付いている場合、その木は順序木とよばれる。

近年のグラフマイニングの成果としては、以下のものが挙げられる：猪口・鷲尾らのグループは、一般のラベル付有向グラフの集合から頻出するラベル付順序木パターンを取り出す手法を提案した。しかし、この手法は、対象を一般の有向グラフとするために、パターン保持にメモリを多く(具体的には $O(n^2)$)利用するため、入力するデータ量が大きくなると高い性能の計算機が必要となる。一方、有村らのグループは対象をラベル付順序木の集合に絞り込むことによって、高速にラベル付順序木の頻出パターンを計算することが出来る。この手法は非常に高速で使用するメモリ量も少なく、大規模データにも適用可能であるが、対象となるデータの種類がラベル付順序木というかなり限られたものになってしまう。このように、この分野は高速かつ対象が出来るだけ広いマイニング手法を模索して、沢山の研究が行われている。

ここで、ラベル付順序木のデータマイニングの応用例として、係り受け木のマイニングを取り上げる。係り受け木とは日本語の文の係り受け構造をラベル付順序木で表したものである。一つの文が一つのラベル付順序木に対応する。また、木の各頂点は文に含まれる文節が対応する。係り受け木はラベル付順序木であるため、上で述べたラベル付順序木に対するデータマイニング手法を用いることができる。図３．２．１は係り受け木の集合とその頻出パターンの例である。この例では、「商品Ａが高いというクレーム」が多いという仮説と、「客１」が頻出するという仮説が得られたことになる。この例でも分かるように、グラフマイニングにおいては、全体におけるパターンの出現数とともに、パターンの大きさが重要であり、複数の頻出パターンが得られた場合、それらをどのようにランク付けするかという点も考慮する必要がある。たとえ出現数が十分に多くてもあまりに小さなパターンは情報量が少なく、知識として有効でない場合があるからである。この係り受け木集合の頻出パターン抽出によるデータマイニングシステムは、コールセンターのクレーム解析に実際に使われた例があり、その有効性も確かめられている。

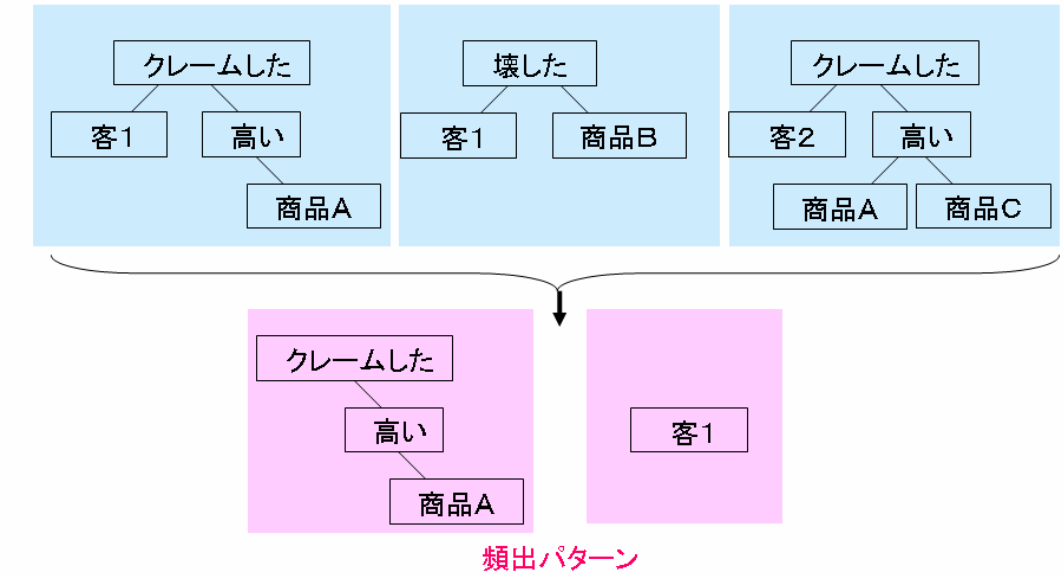

図３．２．１　係り受け木の頻出パターン抽出

４．Web 2.0 について

４．１　Web 2.0 とは

Web 2.0 とは、Tim O'reilly 氏が生み出した概念であり、従来の WWW における静的なサービスに対し、次世代にあるべき新しいウェブのあり方に関する総称である。Web 2.0 という言葉は、言い回しから連想されるような厳密な規格やフレームワークを指すのではなく、おおまかなサービスのあり方の方向性を指しており、その概念の捉え方も人によって様々である。しかしながら、Web 2.0 という言葉について共通に持たれた概念もたしかにあり、また、Web 2.0 に含まれるとされているサービスの間にはある一定の共通する特徴があることも事実である。

O'reilly 氏は自身のWWWで「What is Web 2.0」という論文を発表し(*)、その概念を規定しようとしている。

[* http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html]

この論文によれば、まず、従来の WWW のサービス(これを O'reilly 氏は Web 1.0 と呼んでいる)に対して、それぞれ対応する Web 2.0 の概念を例示することで、Web 2.0 の概念について明確にしようと試みている。以下、その O'reilly 氏の論文に掲載されていた例を引用する。

Web 1.0　　Web 2.0

DoubleClick --> Google AdSense

Ofoto --> Flickr

Akamai --> BitTorrent

mp3.com --> Napster

Britannica Online --> Wikipedia

personal websites --> blogging

evite --> upcoming.org and EVDB

domain name speculation --> search engine optimization

page views --> cost per click

screen scraping --> web services

publishing --> participation

content management systems --> wikis

directories (taxonomy) --> tagging ("folksonomy")

stickiness --> syndication

これらの例のうち、いくつか Web 2.0 を特に特徴つけるものについて詳しく述べる。

DoubleClick 社は Web service を用いた WWW 広告を事業としたという点において先駆的な会社であり、最盛時には 2000 社の広告収入を得ることに成功した。しかしながら、広告の契約は正式の販売契約を通してのみ提供されたため、ある程度以上の規模の会社としか契約に至らなかった。一方、Google AdSense は契約を自動化し、大量の小規模の会社と契約することで、市場を大きく広げることに成功した。このように、一つ一つは小さい量

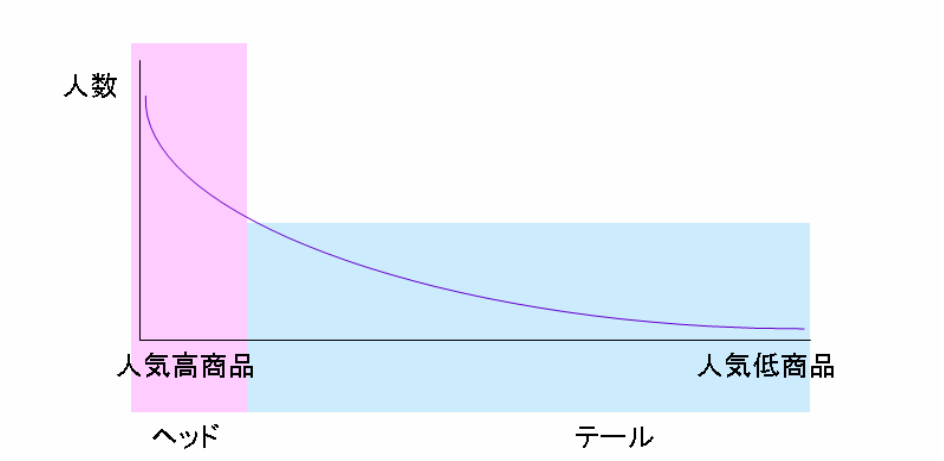

図４．１．１　ロングテール概念図

のものであっても世界中からすべて積み上げれば全体として非常に大きな量になりうるという考え方を「ロングテール」とよぶ（図４．１．１参照)。この「ロングテール」を重視

したサービスは Google, Amazon など、Web 2.0 で成功したとされている会社に共通のものである。

「Wikipedia」は「集合知」の成功例として知られている。Wikipedia には誰でも記事を投稿し、編集することができる。Wikipedia の日本語版のトップページでは、誰でも記事を投稿・編集できることが明記されている。誰でも記事を書けることから、間違いを含んだ記事が投稿されることも当然有り得るが、誰でも記事を編集できるため、誰かが間違いを見つけ出し、誰かが訂正してくれるという、いわば性善説に基づいたシステムである。一見、間違いが容易に入り込めそうに感じられるが、Wikipedia の間違いの割合は市販されている百科事典と同程度か少ないくらいであるという報告もある。この「特定の専門家による知」vs「集合知」の考え方は、ソフトウエアの開発における「伽藍モデル」vs「バザールモデル」の議論にも似ており興味深い。O'reilly 氏は先に紹介した論文「What is Web 2.0」の中で、「Web 2.0 の本質は集合知を利用して WWW を地球規模の脳に変えることだ」とも述べている。集合知が Web 2.0 の概念のなかで最も重要な考え方の一つであることはこのことからも分かる。

ディレクトリ型の整理(taxonomy)からタグ付けによる整理(folksonomy)への移行も Web 2.0 の大きな特徴の一つである。さらに沢山の人々のタグ付けの結果を収集し解析することで、情報を分類解析することもできる。これは一種の集合知であり、ソーシャルブックマークサービスはその代表的なものである。

個人ホームページと対照されているブログも Web 2.0 の特徴をよく表したサービスであるといえよう。ブログは日記の体裁をしたWWW上のサービスであるが、単なるWWWの日記サービスであれば Web2.0 以前より存在していた。ブログの最も大きな特徴は RSS(RDF Site Summary)と呼ばれる技術にある。RSS は、ウェブサイトの更新情報を簡単にまとめ、配信するための文書フォーマットである。ブログサービスでは RSS が使われているため、ブログを更新するたび、そのブログを購読している人々は更新通知を受け取ることができる。RSS はその特性からブログの他、ニュース配信などでも使われている。現在最も使われている RSS は RSS1.0 であり、XML で記述され、URI、タイトル、そのページに対する短い説明を書くことが出来る。

４．２　Web 2.0　とデータベース

Web 2.0 を特徴つける概念の一つとしてロングテールや集合知や folksonomy を挙げたが、これらの概念に共通することは、いかにデータを効率よく集め、それらをうまく管理するかにサービスの質が依存していることである。ロングテールは小さなニーズを大量にかき集める必要があるし、集合知はそのサービスを沢山の人に利用してもらわなければ、仕組み自体がうまく働かない。具体的な Web2.0 企業を見渡してみても、Google は自社内に

WWW コンテンツという大量のデータを溜め、Amazon は商品の評価結果や売り上げデータを大量に溜めることでサービスをうまく動かしている。

そして、Web2.0 が最終的に目指すものが WWW を一つの脳とすることであれば、ただデータを漫然と溜めることでは実現できるはずがない。適切にメタデータを付け、それらのデータを他者と共有できる形で溜めていくことで、より Web 2.0 が目指す形でのデータベースができていくものと期待される。

５．グリッドコンピューティングについて

グリッドコンピューティングとは、ネットワーク通信を通して、複数のコンピュータ資源をひとつの仮想的コンピュータシステムとして利用するための仕組みである。元々は遊休計算機資源を有効に活用するために作られた仕組みだったが、現在は、大規模計算を効率よく行うための仕組みとして利用されている。さらに近年では、大規模なデータを扱う必要のある科学研究分野で「データ・グリッド」が提唱・開発され、大量のデータを保存・利用するための手段として利用されている。

グリッドは仮想的に一台の計算機であるかのように振舞う。「グリッド」の言葉は本来電力網を意味する。家庭で電気を使う際、われわれはその電気がどこの発電所で作られ、どのようなルートで供給されるか全く気にする必要がない。グリッドコンピューティングについても、エンドユーザはグリッド内部で、どのようなマシンがつながっており、どのようにデータが交換され、あるいは計算が配分されているかを気にする必要がない。しかし、このグリッドコンピューティングの目指す環境を実現するためには、利用する計算機が異なる組織に属していたとしても、それらの計算機のプラットホームがばらばらであっても動的に連携できる仕組みが必要である。

グリッドに属する計算機が異なる組織に属する場合、計算機の利用ポリシーが組織によって様々であるため、それが潤滑な計算機資源提供をさまたげる可能性がある。安定した計算資源をグリッドに供給し、ユーザがグリッドを安心して使用できるためには、この異なる組織間をまたいで、共通ルールを定めることが必要となる。共通ルールに従う複数の組織を仮想組織(Virtual Organization)と呼ぶ。仮想組織の概念は 2001 年に Globus（*）によって提唱された。

[Globus とは 1995 年に設立された、グリッドに関する基礎的な技術の開発をすすめる団体である。米国からはアルゴンヌ国立研究所、南カリフォルニア大学情報科学研究所、シカゴ大学が、ヨーロッパからはスコットランドのエディンバラ大学とスウェーデン並列コンピュータセンターが参加している。]

さらに、通信技術としてのハードルとして、グリッド間の通信プロトコルの標準化と通信管理方法の問題がある。グリッド環境を実現させるためには、複数の計算機をプラットホームや OS の壁を超えて連携させる必要があるからである。かつて、グリッド間通信におい

ては、独自形式の通信プロトコルが使われてきた。しかし、2001 年に Globus がグリッド間通信に Web service 技術を使うことを提唱して以来、グリッドの通信には Web service 技術が使われている。以前の独自形式の通信プロトコルが、大規模計算のための計算機パワーの振り分けに特化していたことに対して、Web service による連携はメッセージ交換というゆるやかな結合であるため柔軟性に富み、計算機パワーの振り分けからデータの共有まで様々な用途に利用できることになった。事実、最新版の Globus Toolkit (Globus が提供する科学系グリッドの標準となっているミドルウエア、現在の最新バージョンは 4.0.4) は WSRF (Web Service Resource Framework) とよばれる Web service のためのフレームワークに準拠して実装されており、グリッド資源管理プロトコルやデータ移動のための二次記憶への広域アクセスなど、様々な機能を提供している。

さらに、Web service の導入から一歩進めて、グリッド上で SOA (Service Oriented Architecture)を実現する流れも出てきている。これは科学技術計算方面からの要請というより、むしろ産業界からのニーズに応える形で動いてきている。企業をターゲットとしたグリッドはビジネスグリッドとよばれている。科学技術計算のためのグリッドとビジネスグリッドは、決してばらばらの方向へ進んでいくものではなく、要素技術や標準化の面で共通するものも多い。ビジネスグリッドの発展から科学技術計算のグリッド技術が得ることも多くあると考えられる。

６．データベース統合へ向けた情報技術利用の可能性について

この節では、これまでに調査した情報技術、すなわち、検索システム、データマイニング、Web2.0、グリッドコンピューティングについて、生物情報データベースの統合に応用する場合、どのような可能性があり、またどのような問題点があるかについて議論する。

まず、検索システムについてだが、多種多様かつ膨大な量の生物情報を検索するには、従来技術のキーワード検索や項目による検索ではすでに限界がきており、検索エンジン自体に高度な解析機能、可視化技術が必要となってきている。種類も記述方法も異なる生物情報について、どのように情報を選び、さらに、それをユーザが見やすい形で見せていくかという研究は今後ますます必要になってくるであろう。また UIMA のようなツールの標準化技術は、次々と新しい種類・形式のデータが現れる生物情報においては非常に有効である。なぜならば、始めからすべてのデータの種類を規定して検索ツールを用意しておく必要はなく、新しい種類のデータが現れた際に、そのデータへの検索ツールを作成してプラグインすればよいからである。

データマイニング技術は、人の目で追っていては到底届かない膨大な生物情報を扱うための重要な道具として、これまでも重要な役割を担ってきた。今後もその重要性は変わらないと思われるが、これまで生物情報の中心であった、生物分子情報や文献情報だけでなく、WWWなどに眠る画像や様々な非構造化データをも統合に取り入れていくとすると、それ

らに対するマイニング技術も必要になってくるであろう。また、現在でも膨大な量である生物情報だが、今後、生データがさらに莫大な量になれば、マイニングによって得られる仮説の数も増え、仮説を評価すべきユーザの負担が増大する。したがって、マイニング結果の評価を支援する可視化ツールやマイニング結果をさらにマイニングするようなツールも必要となってくるであろう。

Web 2.0 は第 4 節でも議論したように、「いかにデータを効率よく集め、それらをうまく管理するか」という技術である側面を持っている。その代表的な概念が「集合知」である。生物情報の統合に集合知を利用することができれば、非常に有益な統合データが得られる可能性がある。しかしながら、集合知を利用するためには、多くの人々にそのサイトを知ってもらう必要があり、さらに、参加したくなるための仕掛けも必要である。また、多くの人々が参加してくれる場合でも、そこで得られた情報をいかに管理していくかの仕組みも必要となる。

グリッドコンピューティングの技術は、かつての科学技術計算中心のものから、サービス中心のものへ移行してきている。それに伴い、データ交換方法についても標準化されてきており、その手法はデータベースの統合にも生かせる可能性がある。しかしながら、グリッドコンピューティングのデータ交換は Web service を基本としたものである以上、Web service のもつ特性から離れられず、それにともなった様々な問題をかかえている。たとえば、生物情報データベースに特有の Web service の問題点として、同じ遺伝子、あるいは同じタンパク質に異なった ID がつけてある場合、ユーザが求める結果が得られない可能性がある。これは、Web service ではシンタックスのみを扱い、セマンティクスを扱う枠組みがないことに由来している。したがって、グリッドコンピューティングに限らず、Web service をベースとした統合を目指すのであれば、セマンティクスをうまく扱う仕組み、たとえば Semantic Web などの技術を取り入れていく必要があるだろう。