特集「アクティブマイニング」にあたって

津本 周作*1，山口 高平*2，沼尾 正行*3，元田 浩*3
（*1 島根大学医学部医学科，*2 慶應義塾大学理工学部，*3 大阪大学産業科学研究所）

1. アクティブマイニングとは？

データマイニングは，大量データ中に潜む有用な知識の発見を可能にする，ネットワーク社会の根幹となる技術であり，ナレッジマネジメントの強力な武器になる．しかし，大量データの氾濫による今日の情報洪水の状況下では，1）膨大な情報空間の的確な監視および効率的な情報収集，2）多様な形態の情報源からの価値ある知識の発掘，3）ユーザの視点の変化や状況変化に即応した知識の頻繁な更新の何れもが「情報洪水の起こる以前には直面しなかった」大きな課題を抱えている．このため，情報収集・データ解析・目的設定変更のサイクルがうまく機能せず，空洞化し，個人も組織も情報洪水の中で疲弊しているのが現状である．

これらの問題点に焦点を当て，我々はデータマイニングにおけるデータ解析のサイクルに専門家からのフィードバックを組み込んだアクティブマイニングのフレームワークを提唱した．それに基づき，共通医療データを核として各計画研究が連携して要素技術を開発し，大量データからの知識発見の実用化を目指した“知識発見のらせんモデル”を実践してきた．その結果，現場の専門医からも新しい診療知識発見技術として厚い信頼を得るに至った．

2. アクティブマイニングプロジェクトの目的

領域が内蔵する，上記の三つの課題に対応し，1）自律的に必要な情報源を探索し前処理を実施するアクティブ情報収集，2）種々の構造をもつデータに適した柔軟なマイニングを実現するユーザ指向アクティブマイニング，3）理解しやすい表示と結果に対するユーザの積極的なフィードバック環境を提供するアクティブユーザリアクションの研究を実施し，三つの連携機能を実現する環境を構築することを領域全体の設定目的に掲げている（図1）．領域全体で連携してこれらの課題を解決し実証するため，具体的な問題として，医療と化学薬品データのマイニングを取り上げた．前者に関しては千葉大学医学部より提

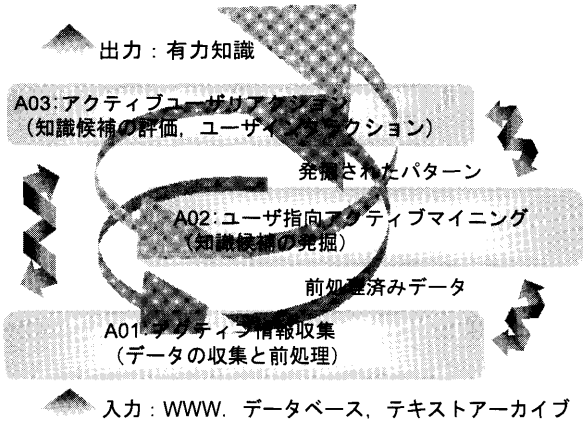

図1 アクティブマイニングプロジェクト（らせんモデル）

供される肝炎疾患データを共通医療データとして各計画研究が取り上げ，上記の三つの機能を統合した“知識発見のらせんモデル”による相乗効果の実証を試みた．その過程で，血液検査データから肝炎の病理像（線維化の程度）を推定する知識を獲得し，侵襲度が高く患者に苦痛を与える肝生検に代わる検査指針の可能性を探る．同時に，21世紀の医療の指針として注目されているEvidence Based Medicine（EBM）（「エビデンスに基づいた医療」）の実践に寄与する有効な手段であるかどうかを検討した．化学薬品に関しては，毎年新しく開発される多数の薬品の生理活性に対する部分化学構造から，すでに市場に出回っている薬品に対し必要な警告を与える可能性を検討する．

3. アクティブマイニングプロジェクトの研究組織

本プロジェクトは科学研究費補助金特定領域研究「情報洪水時代におけるアクティブマイニングの実現」（研究期間：平成13年9月～平成17年3月，領域代表者大阪大学産業科学研究所教授 元田 浩）として採択され，この特定領域研究をベースとして研究が進められてきた．研究組織は研究の評価と推進をはかる総括班のもとに，上記目的に記した三つの班を構成した．各班は，それぞれ次のような計10個の計画研究からなる．

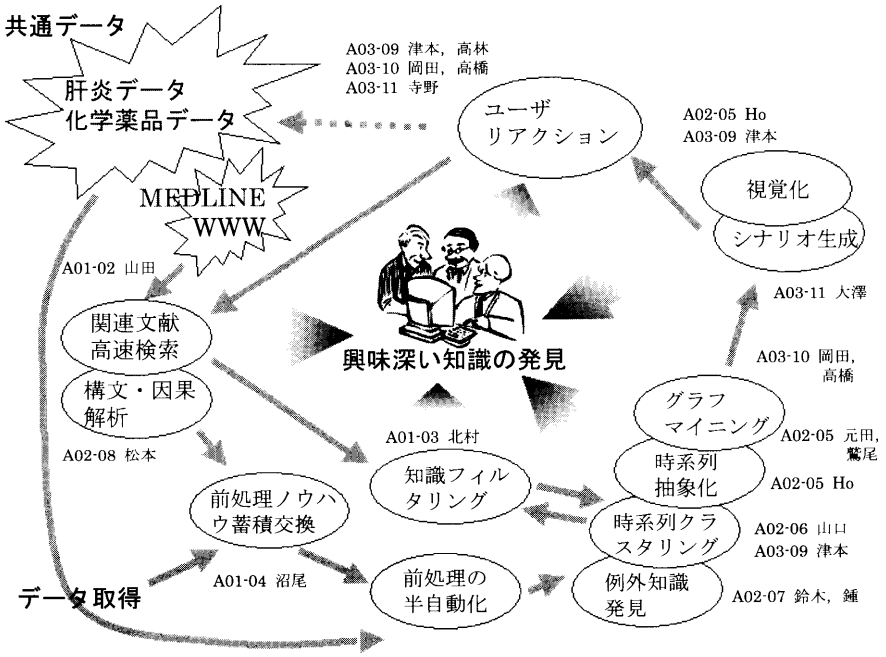

図2 各研究グループの連携図

(1) A01班：アクティブ情報収集

●A01-02: 山田誠二： WWWにおけるメタ情報源の獲得

● A01-03: 北村泰彦：分散動的情報源からのアクティブ情報収集

●A01-04: 沼尾正行：多段階学習方式によるデータ収集と前処理の自動化

(2) A02班：ユーザ指向アクティブマイニング

●A02-05: 元田 浩：構造データからのアクティブマイニング

●A02-06: 山口高平：メタ学習機構に基づくアクティブマイニング

●A02-07: 鈴木英之進：例外性発見に基づくスパイラル的アクティブマイニング

●A02-08: 松本裕治：利用者からの要求を考慮したテキストデータからの知識抽出

(3) A03班：アクティブユーザリアクション

●A03-09: 津本周作：ラフ集合に基づくアクティブマイニングによる診療情報生成システムの開発

●A03-10: 岡田 孝：アクティブマイニングによる化学物質群からのリスク分子発見

●A03-11: 大澤幸生：ヒューマンシステムインタラクションに基づく知識の評価と選択

初期の目標を達成するためには，各計画研究間の密接な連携と領域専門家の協力が不可欠である．専門家を含めた会議を定期的に開催し，各計画研究の進捗状況を全員が把握すると同時に，問題点を徹底的に討議し解決策を全員で議論してきた．特に，肝炎データ解析にはA03-10を除くすべての計画研究が参加しており，これらの会議にはデータ提供者の千葉大学医学部の専門医にも参加してもらい，解析結果の評価を行った．このプロジェクトの間，専門家を含めた会議は48回にも及び，これらの医学・化学に関するデータ解析以外にも，多くの研究テーマが並行して推進され，それらの成果を交換しながら，領域全体としての研究成果に集約されていった．共通データ解析の目的は，各要素技術を連携して前処理，マイニング，評価のサイクルを繰り返し，“科学発見のらせんモデル”を実践・実証することであるが，これが地理的にも離れた計画研究を連携させる大きな求心力になってきた．総体的には各計画研究は所属する班の目標達成のための研究遂行に力を注いできたが，図2に，これらの各研究班の関係を示すように，互いに少しずつオーバラップしながらマイニングの全プロセスを経験することができた．

4. 主な研究成果

各研究班においては，以下の成果が得られている．1) アクティブ情報収集では，分散情報源からの効率的な情報収集，関連情報による発見知識のフィルタリング，前処理手順の自動獲得，効率的な情報収集のためのメタ情報源の自動学習，などで，2) ユーザ指向アクティブマイニングでは，ユーザの使用目的に合致したマイニングアルゴリズムの自動構築，時系列データのクラスタリングと可視化，時系列データの抽象化，時系列データの決定木学習，専門家が容易に関与し得る環境の構築，構造データからの共起パターンの高速発見，スパイラル的例外性発見などで，3) アクティブユーザリアクションでは，元データやマイニング結果の知識の視覚化とクラスタリング，視覚化を通した専門家の主観的発見プロセスのモデル化，専門医が直接マイニングに関与できるインタフェース，デ

ータからのシナリオ生成などで，それぞれ顕著な成果が出た．共通医療データである肝炎データの解析では，2年目から各要素技術を連携して前処理，マイニング，評価のサイクルが回り始めた．専門医とのインタラクションも活発に行われており，専門医の興味を引く具体的な成果も出ている．化学薬品のデータマイニングについても，2年目にドーパミン拮抗薬の類似構造検索で高血圧治療剤が発見され，高血圧治療に伴う精神面への副作用の恐れを提示でき，さらに知識表現の洗練および分類精度の向上に成功し，対象データ範囲を拡張して，目標とするリスク警告システムを稼働させる見通しを得た．

5. 本特集のあらまし

本特集では，アクティブマイニングプロジェクトで得られたこれらの研究成果について詳しく紹介している．研究成果はそれぞれ盛りだくさんであるため，ページ数の都合で，すべてを紹介できていない場合もあることを了承していただきたい．

以下，各解説に示されている主要成果を要約する．

(1) アクティブ情報収集

WWWにおけるメタ情報源の獲得：対話的文書検索，WWWからの情報ストリームの獲得，Webコミュニティの発見と獲得，を要素技術とし，それらの統合により，効率的かつ効果的なメタ情報源の獲得を実現した．(1) 対話的文書検索：従来の適合フィードバックに対し，関係学習とサポートベクタマシンを適用して，さらなる検索性能の向上をはかった．(2) WWWからの情報ストリームの獲得：キーワードを抗体，文書を抗原として免疫ネットワークを構成し，免疫系の特性を有効に活用した情報ストリームの獲得を実現した．(3) Webコミュニティの発見と獲得：Webページにおけるリンクの共起を検索エンジンを使って評価することで，密に連結しているWebページの塊であるWebコミュニティを発見，獲得する方法を開発した．

分散動的情報源からのアクティブ情報収集：分散動的情報源からの効率的な情報収集法を開発し，発見が新規なものであるか，既知のものであるかを判定するフィルタリング手法を提案した．さらに検索された関連ルールをクラスタリングし，順位づける方式を考案し評価した．アクティブマイニングプロジェクトの共通領域である肝炎データマイニングへの応用として，医学生物学文献データベースMEDLINEからの情報収集に基づく発見ルールフィルタリングシステムを開発した．

多段階学習方式によるデータ収集と前処理の自動化：前処理のプランニングを行うシステムを矢田らが開発しているオープンソースプラットフォームMUSASHI上に実現した．マイニングを効果的に進めるためのデータの前処理として，属性の重み付けによるマイニング過程の制御（文献データベースに基づいて属性を選択し，重み付ける）および実例の重み付けによるマイニング過程の制御（実例の分布に基づいて，実例を選択し，重み付ける）について検討した．時系列データの前処理として，同一値を取る期間に基づく前処理間隔不定な時系列データを内挿する前処理により，マイニング結果を改善した．さらに，専門家，データ解析者との間の情報収集を高めるため，伝言ゲーム型の情報収集の方法を考案した．

(2) ユーザ指向アクティブマイニング

構造データからのアクティブマイニング：構造の大きさにほぼ線形な処理時間で多頻度連結部分グラフを抽出するアルゴリズムを開発し，肝炎データに適用し，専門医が興味を示す結果を得た．さらに，グラフ抽出アルゴリズムを拡張し，分類に効果的な属性を逐次的に構築・利用するグラフ構造データ向きの決定木生成法を提案し，知識を抽出した．また，規則やデータの階層構造を視覚化し，ユーザがマイニングプロセスに積極的に関与し必要なモデルを選択できるマイニング環境を構築し，時系列データ抽象化による肝炎データ解析法を開発し，B型肝炎とC型肝炎の違い，生検データに基づく肝炎の進行状態，インターフェロンの効果の予測ルールを発見した．さらに，これらの手法を統合するオープンソースプラットフォームMUSASHIを開発した．

メタ学習機構に基づくアクティブマイニング：メソッドや評価基準のプリミティブ群（リポジトリ）から適切なアルゴリズムを構成する「構成的メタ学習」を提唱し，これらのリポジトリを組み合わせて，ユーザの使用目的に合致したマイニングアプリケーションを半自動合成するシステムを開発した．本研究領域の共通データである慢性肝炎データセットに適用したところ，ある条件下でGPTが約3年周期で変動するという，GPTは単調変化するという定説に反するルールが得られた．

例外性発見に基づくスパイラル的アクティブマイニング：既知の例外性をデータ，知識，環境に照らし合わせ，新しい例外知識を発見する手法を開発し，肝炎データに適用し医師の興味を引く例外性を発見した．さらにより興味深い例外性を発見するために，確率モデルを用いて，病院検査データを類型化し，可視化する手法と動的伸縮法による時系列属性をもつ決定木の学習法を提案した．さらに，この技術を洗練し，慢性肝炎患者データから特定される例外患者および全体的傾向に関する発見知識を検討・再評価した．

利用者からの要求を考慮したテキストデータからの知識抽出：未知語を含む英文文書内の単語の品詞推定，文中の基本句の自動抽出および係り受け解析，テキストからの専門用語の抽出と分類のための基本的な手法を確立し，これらをベースに，大規模医学文献データベースMedLine文書の構造解析を実施し，背景説明や結論部の同定に成功した．さらに，因果関係の自動抽出の検討に着手した．

(3) アクティブユーザリアクション

ラフ集合に基づくアクティブマイニングによる診療情報生成システムの開発：時系列データを平滑化・セグメント化し，多重スケールマッチングを用い，時間スケールの違う検査値推移パターンを比較可能とした．さらに，ラフ集合論の識別不能性の概念に基づき，セグメント化した系列を分類するクラスタリング法を開発した．肝炎データに適用し，GPTの時系列変化とC型肝炎のインターフェロンによる投与効果の強い相関関係，血小板数が長期予後予測指標として利用可能，血小板数から線維化や活動度が予測可能など多くの知見を得た．さらに検査間隔の分散に着目した不均質系列の判別法を開発し，マイニング結果のスパイラル的活用を試みた．

アクティブマイニングによる化学物質群からのリスク分子発見：芳香族ニトロ化合物の変異原性を解析し，ortho位置換基の立体障害が重要因子となることを発見し，カスケードモデルの有用性を確認した．同一手法による解析で，発ガン性予測の国際コンテストで参加14グループ中，第1位の評価を得た．さらに，ルールを組織化し，データの特徴を多面的に捉えることを可能とし，ドーパミンアンタゴニスト活性化合物に適用し，各受容体に対して専門家が納得する特徴的な部分構造群を発見した．並行して，フラグメントを質量数で特徴づけたTFSの類似性から活性分子を同定する手法を開発し，構造類似性検索を可能にし，異なる活性クラスに属するにもかかわらず構造類似性の極めて高い例外分子が見いだされることを示し，リスクレポートへの有効性を実証した．本手法を3次元化し，ドーパミンアゴニスト活性を有する化合物群の3次元特徴フラグメントを抽出した．

ヒューマンシステムインタラクションに基づく知識評価と選択：チャンス発見の二重らせんモデルを提案し，それに基づきキーグラフを仲介にしたヒューマンマシンインタラクションを促進するシステムを開発した．医療データに適用し，肝炎進行シナリオ生成を試み，B型肝炎では進行中の黄疸が肝硬変に至る危険性をもつこと，C型肝炎では，インターフェロンはウイルス抗体が生成される初期の投与が効果をあげることなどの知見を得た．さらに，鉄代謝がB型肝炎における肝硬変からの回復，C型肝炎におけるインターフェロンの効果を決定づけるうえで特に重要性が高く，鉄代謝を支えるタンパク質が肝炎治療に有益であるとの予想を示す結果を得，専門医から高く評価された．

6. アクティブマイニングの今後

プロジェクトでは，参加者全員が問題の難しさと研究の進展の喜びを共有し，計画研究間の連携による知識発見のらせんモデルの実践がスムーズに進展してきた．ただし，得られた知識の評価は専門医に頼らざるを得ず，専門医の負担が増大していった．しかし，医療分野の専門家も，徐々にアクティブマイニングの魅力に取りつかれ，知識の評価，助言に積極的であった．総括班が非常にうまく機能し，全員の参画意識向上に寄与した．本プロジェクトの成功の要因は，共通データ解析が各研究班間の連携，班内の各計画研究間の連携の大きな求心力になったことであろう．本研究は，単に各研究班の研究成果のみならず，専門家を含めたアクティブマイニングのプロセスを適用することができ，いわゆるFayyadの提唱したKDDプロセスをより精緻な形で検証できたことにある．今後，これらの成果がデータマイニングの研究をさらに一歩進める糸口となることを期待したい．