情報洪水時代における
アクティブマイニングの実現

領域番号 759

平成13年度科学研究費補助金　特定領域研究（B）

研究成果報告書

平成14年3月

領域代表者　元田　浩
（大阪大学産業科学研究所・教授）

はじめに

特定領域研究（B）「情報洪水時代におけるアクティブマイニングの実現」（略称：アクティブマイニング）は，本年度（平成１３年度）から平成１６年度までの，４年間の研究としてスタートしたものです．

周知の通り，通信技術を含む計算機ハードウェアの急速な進歩により，大量情報が各種ネットワークを流通する時代に突入しました．この状況は，情報洪水にたとえることができます．実際，１）膨大な情報空間のどこを見ればよいのかが分からない，２）見る場所が同定できても，その中から目的にかなった価値ある知識を簡単には取り出せない，３）状況変化に即応できず，頻繁な知識の更新に対応できない，などの大きな問題がクローズアップされています．情報収集・データ解析・目的設定変更のサイクルが高速回転し，個人も組織も情報洪水の中で疲弊しているのが現状です．

このような状況を打破するために，新しいマイニングの枠組み「アクティブマイニング技術」を実現することを目的として，本特定領域研究がスタートしました．アクティブの名が示す通り，システム側からの情報源への積極的な働きかけ，目的に合致した質の高い知識の効率的な発掘と効果的な提示，ユーザ側からのシステム側への迅速なフィードバックの実現を標榜しております．

この特定領域研究には我国を代表するデータマイニングの研究者が多数参加し，上記した３つの課題に対応して，以下に示す３つの研究班を構成し，相互の有機的な連携を取ながらマイニングのための方法論を研究開発しています．

（A01 班）アクティブ情報収集

不特定・非定常・大規模・分散知識源の中から，ユーザの目的や興味に合致するデータやそれらの関連を効率良く探索し前処理するための情報収集技術を，メタ情報源の活用，ヒューリスティック探索知識の活用，機械学習法の活用など，最新のＩＴ技術を駆使して開発する．

（A02 班）ユーザ指向アクティブマイニング

多様な形式や多種の情報源に対応できる，汎用性と状況の変化に対応できる柔軟性を持つマイニング手法を開発する．とくに，テキスト情報に代表される半構造化データ，巨大分子化学情報・ネットワーク情報に代表される構造化データからのマイニング，これら個別のデータに最適なマイニング手法の自動構築，状況変化検知に強力な例外性の発見技術に注力する．

（A03 班）アクティブユーザリアクション

具体的な問題領域（医療，化学薬品）を対象にマイニングシステムを構築し，発掘した知識を，ユーザにとって有用なものとするための仕組（知識の表示法，評価手法，ユーザからの効果的なフィードバックの手法）を具体化する．

このような目標の下に，これらの３つの機能を統合して得られる相乗効果により，知の上

昇スパイラルを実現し，情報洪水から人々を救出する有力な手段を提供したいと考えております．

本報告書は，こうした特定領域研究（B）「アクティブマイニング」における初年度の研究成果と研究活動をまとめたものです．最初に，全体の企画・調整・評価を主な任務とする総括班の活動と今年度の成果の概要，来年度の計画をまとめ，以下，各班毎に計画研究の成果を報告します．なお，総括班としての活動の詳細は本特定領域研究のホームページ

<http://www.ar.sanken.osaka-u.ac.jp/activemining/>

にありますので，そちらもご参照下さい．

最後に，本特定領域研究を支えて下さった関係各位に厚くお礼を申し上げますと共に，本研究のいっそうの発展のために，今後とも変わらぬご指導ご支援を賜りますようお願い申し上げます．

平成 14 年 3 月
特定領域研究「アクティブマイニング」
領域代表者　元田　浩

目次

総括班

総括班 活動報告

領域代表者　元田 浩　(大阪大学産業科学研究所)

1　研究目的と研究計画

特定領域研究 (B)「情報洪水時代におけるアクティブマイニングの実現」の総括班として，領域全体の企画・調整・評価を行い，本特定領域研究が所期の目的を達成するように研究活動を総括するのが目的である．具体的には，以下の研究活動を計画し実施した．

1)　総括班会議の開催

2)　代表者会議の開催

3)　共通医療データ解析に対する取組み方針の策定

4)　セミナーの開催

5)　班会議の開催と中間成果の発表

6)　研究会の企画と成果の発表

7)　ホームページの開設と運用

8)　アクティブマイニング特集号の企画

9)　出版の企画

10)　国際ワークショップの企画

11)　研究成果発表会と報告書の刊行

1.1　総括班会議と代表者会議

領域全体の企画・調整・評価等を行なうための総括班会議を 2 回開催した．第 1 回会議は，平成 13 年 11 月 12 日にはこだて未来大学 5 階会議室で開催し，本特定領域研究の今年度の全体計画，共通医療データと各班の取組み，各計画研究の今年度の計画と協力体制を議論し，来年度の計画 (国際ワークショップ，研究会など) を検討した．第 2 回会議は，平成 14 年 3 月 6 日に筑波大学ビジネス科学研究科大学院会議室で開催し，本年度の活動を総括し，来年度の活動計画を策定した．

具体的な研究遂行上の諸問題，各計画研究間にまたがる技術的な細部を議論するため代表者会議を 3 回実施した．第 1 回会議は，平成 13 年 7 月 21 日に東工大百年記念会館で開催した．本会議が実質的な本特定領域のスタートで，本特定領域研究発足までの経緯を説明し，各計画研究の研究計画と目標を議論し，推進体制を策定した．第 2，3 回会議は第 12 回ヨー

ロッパ機械学習国際会議と第5回ヨーロッパ知識発見とデータマイニング国際会議に参加した代表者で，平成13年9月4日，6日にドイツのFreiburg大学にて開催した．この会議で，本特定領域研究で各計画研究が共通に使用するデータを議論し，千葉大学医学部附属病院医療情報部高林助教授より御提供頂ける肝炎に関する医療データを使用することを決定した．

1.2 共通医療データ解析に対する取組み方針の策定

本特定領域研究では（A01班）：アクティブ情報収集，（A02班）：ユーザ指向アクティブマイニング，（A03班）：アクティブユーザリアクションが単独で研究を遂行するのではなく，各班が相互に連携をとり，これらの３つの班の成果を統合して得られる相乗効果により，知の上昇スパイラルを実現し，情報洪水からの脱却に有力な手段を提供することを目的としている．そのためには，各班(各班に属する計画研究）が共通のデータを扱い知見を共有することが不可欠である．1.1で述べたように，千葉大学医学部より貴重なデータを御提供頂けることになった．医療データの取り扱いには患者が特定され得ないようにデータを加工していただき，データの開示制限，守秘義務を明記した契約書を本特定領域研究代表者と高林助教授で交わし，解析を開始した．

肝炎に関する共通データ

データの概要． 肝炎に関する共通データはB型・C型肝炎に関するデータ，第一内科第二研究室で記録している慢性肝炎フォロー患者の肝生検結果と，それらの患者の血液検査データからなっている．ここから，アクティブマイニングの手法を適用して，どのような知識発見がなされるか検討，さらに新たな知識が発見が出来たときには，その妥当性について検証することを目的として，データが提供された．

本データの対象は千葉大学第一内科で外来フォローしているB型C型の慢性肝炎患者で1982年から2001年までに肝生検を受けた者(計771名) に限り，データ集合は次の二つのデータから構成される．一つは，第一内科から提供されるデータは上記患者基本情報(年齢, 生年月日等), 病理分類, 生検の日時, 生検結果，輸血等の既往歴, インターフェロン投与期間についての情報を含んでいる．もう一つはそれぞれの患者に対応した検体検査データ(血液・尿検査) を病院情報システムから抽出した時系列データである．検体検査は，大学病院内の検査機器で測定可能なもの(院内検査) と外部の検査会社に委託して測定する(外注検査) が含まれる．院内検査は，検査技師のコメントを含めて230種類が，外注検査は，コメントを含めて，753種類含まれており，全体で，983種類の項目によってデータ集合が構成されている．このうち，771名の患者全員が検査を少なくとも1回以上受けていると推定できる検査(つまり総度数が771以上の検査) は検体検査で約76種類，外注検査は3種類である．これらの統計を表1にまとめた．

ウィルス肝炎とは． ウィルス肝炎には主としてA型，B型, C型があり，この中で，慢性化し，肝硬変，肝臓癌に至る可能性があるのはB, C型である．この中で，肝硬変，肝臓癌に至る危険度の指標としては，肝実質の繊維化が考えられ，繊維化が高度に進んだ終末像が肝硬変である．ただし，すべての慢性肝炎の患者が肝硬変に移行するわけではなく，どのような

表 1: データ集合中の項目数

検査種類	総数	コメント	度数 771 以上	度数 4626 以上 (771×6)	度数 9252 以上 (771×12)
院内検査	230	22	76	48	42
外注検査	753	3	3	0	0
計	983	25	79	48	42

機序で肝硬変に至るかはよくわかっていない．唯一，危険因子として指摘されているのは，アルコール摂取量である．ただし，今回のデータでは，アルコール歴に関する情報を含めることができなかった．

上記の B 型，C 型の中で，慢性肝炎→肝硬変→肝細胞癌という典型的な進行パターンをとるのは C 型に多いことが知られており，C 型慢性肝炎の治療にインターフェロンが有効であることが，ここ 20 年の研究で明らかとなってきた．インターフェロンは元来，ウィルスに感染した動物が出すタンパク質であり，ウィルスの増殖を沈静化，駆除する役割を持っている．

C 型肝炎にはいろいろな遺伝子型があり，100%の C 型肝炎ウィルスに効果があるわけではなく，ある種のタイプではインターフェロンの効果があまり期待できないことが知られている．また，インターフェロン投与はインフルエンザ様の症状を副作用として引き起こすことが知られ，この副作用によって投与を中止することもある．

肝炎の進行度等を表す指標として使われている属性. 肝臓が正常な場合，肝臓の 1/10 でも体をまかなえると言われているため，それほど検査値は変化しないと考えられる．このため，肝炎の初期の段階では，値の変化はそれほどない（特に慢性肝炎の場合）．逆に，肝炎の初期の段階で値の変化が大きいならば，肝炎以外の影響が考えられる．

1. ウィルス量：肝炎の活動性を示す (B 型肝炎：HBV-DNA, C 型肝炎：HCV-RNA)
2. 肝逸脱酵素 (GOT, GPT)： 肝炎の活動性，肝臓の障害を示す: 元々，GOT, GPT は細胞内の代謝系酵素であり，これが血中に存在するのは細胞が崩壊した際に，その細胞から流れ出すためであり，この濃度が高いことは肝および心臓をはじめとした筋組織の崩壊の度合いを予想させる．
3. 膠質反応 (TTT, GTT): 免疫グロブリンの量を反映するものであり，炎症が起こっていると値が上昇する．
4. ChE（コリンエステラーゼ），Alb(アルブミン: 肝臓で合成される酵素・蛋白質の量を示す．肝臓が悪化すると，物質が作られにくくなるため値が下がる
5. T-Bil（総ビリルビン），アンモニア (NH_3): ビリルビンは古い赤血球から放出されたヘムを原料として肝臓で合成され，胆汁として小腸に排出される (コレステロール等脂質の吸収を助ける成分)．一般に，ビリルビンは肝細胞由来のものと，この胆汁成分が腸か

ら再吸収されて血中に現れているもの (これは肝細胞に再吸収され，再利用される) で，肝細胞が破壊されると，血中にビリルビン自体が流入するため，血中濃度が上昇する．

6. アンモニア (NH_3): アミノ酸の代謝産物であり，体内では毒性物質である．肝臓で代謝されて尿素として無毒化されるが，肝臓が悪化すると，代謝されなくにくくなる血中濃度が上昇する．

アクティブマイニングへの期待． 慢性 B 型，C 型肝炎がどのような経過をたどって，肝硬変，肝臓癌に至るかについてはよくわかっていない．アクティブなデータ解析によって，次のことについての知見が得られるのでないかと期待される．

1. 病理像と血液検査データとの相関性: 肝炎の病理像（繊維化の程度）と血液検査データとの間にある程度の相関があるのではないかと推測されているが，明らかな知見ではない．肝生検は大変侵襲度の高い検査であり，簡単にできる検査ではないため，できれば，肝生検をせず,（生検）検査以外の血液データからどれくらい疾患の予後が予測できるかに関しては医学的に興味がある．
2. 肝炎の病理像（繊維化の程度）と発ガンまでの期間
3. 血液データと発ガンまでの期間
4. 時系列に関する血液データ積算の有用性
5. B 型肝炎と C 型肝炎の経過の違い: B 型肝炎と C 型肝炎は全く違う異なるウィルスであり，肝硬変，肝臓癌に至る経過が異なることが知られている．実際に B 型と C 型を分けた際に，どれくらい違うかを解明できれば，極めて興味深い．
6. INF 治療の有用性: 統計的データとして，どれくらい効いたか，あるいは再発したかについての知見はあまりない．
7. GOT, GPT が 「進行速度」 の指標となっているが，実際にどうか? 速度 × 時間 = 距離（肝炎の進行度）が成り立つか?

データに関する問題点． 以下の要因で，データの解析が技術的に困難になっている．

1. 病理分類の混在: 新しく提供されるデータにおいては新分類に統一する予定である．
2. 時期により検体検査測定法の違いがある: 数回にわたり測定法の変更があった際，回帰式の扱いをどうしたらよいか同じ尺度での絶対値として扱っていいのか．
3. 病態が悪いことを前提として検査を行っていることがバイアスとなっている可能性がある: 一般の検査ではデータをとらない検査項目があり，その検査に対しては病態の悪い患者さんにしかデータをとらないので，異常値の頻度が多くなってしまう可能性がある．特に，レコード数が少ないものは，病態が悪いことが前提で検査している検査項目であるため，バイアスがかかっている可能性がある．

Evidence Based Medicine (EBM) との関連

共通医療データからのアクティブマイニングは，現在，新しい医療行為として注目されている Evidence Based Medicine (EBM)（「エビデンス（科学的根拠）に基づいた医療」）を実践するための強力な手段を提供する．従って，本特定領域研究はアクティブマイニングの手法を開発するだけでなく，共通データとして医療データを用いた実践を行なうことにより，実際の医療にも貢献できる可能性がある．

EBM とは　患者に対して，医療情報の妥当性・信頼性を十分ふまえた上で，確実で明確な臨床判断を行なうことを重要視する医療方針であり，これを実行するためには，経験と専門知識，現状で利用可能な最も妥当な客観的根拠，これらに基づいた医療を可能にする環境が必要となる．実際の患者から必要な根拠を探し出す方法論から，根拠を評価し，実際の患者に適用すべきかどうか，適用するならどう実際に行なうかを判断する方法まで含む，包括的で実際的な手法であり，アクティブマイニングだけで対処可能なものではない．しかし，患者の医療データや文献データから必要な根拠を探し出すことはまさにアクティブマイニングが目指しているものである．

このような背景を受け，EBM に関するセミナーを平成 13 年 9 月 28 日に大阪大学産業科学研究所にて実施した．講師は本特定領域研究の計画研究代表者の一人でもある津本教授に依頼した．講演の具体的な内容は研究計画 A03-09 に報告する「アクティブマイニングと EBM」を参照して頂きたい．

1.3　班会議の開催と中間成果の発表

各計画研究の進捗状況の把握と班内の計画研究間の調整を図るため，班会議を 2 回開催した．A01 班の第 1 回班会議を平成 13 年 12 月 26 日に東京工業大学大岡山キャンパス西 8 号館 E 棟 5 階コラボレーションルームで開催し，アクティブ情報収集の医療情報に対する取り組みの現在までの成果を報告した．A01 班は千葉大学医学部から後提供頂いたデータを直接解析するというよりは，これに関連した情報を効率よく収集することが目的なので，入手可能な EBM に関する公開医療情報の調査結果や，その中から目的のものを同定する手法に関して議論が白熱した．A02 班，A03 班の第 1 回班会議を合同で平成 14 年 1 月 30，31 日に千葉大学医学部附属病院会議室にて開催した．A02 班，A03 班はデータを直接解析するので，御提供頂いた肝炎に関するデータに対する事前解析の結果を発表し，高林助教授や横井医師からコメントを頂き，また，幾つかの疑問点に答えて頂いた．

1.4　研究会の企画と成果の発表

平成 13 年 11 月 12 〜 14 日に人工知能学会の第 46 回基礎論研究会， 第 54 回知識ベースシステム研究会合同研究会にてアクティブマイニングの特集を企画し，本特定領域の全計画研究から成果を発表した．本特定領域研究以外の発表も含め，総計 42 件もの発表があり，アクティブマイニングに対する関心の高さを知ることができた．次年度の研究会として，情報処理学会の第 128 回知能と複雑系研究会，人工知能学会の第 56 回知識ベースシステム研究会合同研究会にてアクティブマイニングの特集を企画した．平成 14 年 5 月 23，24 日に韓国釜

山市の韓國海洋大學校にて開催する．この研究会の成果は平成 14 年 5 月 28 〜 31 日に開催される第 16 回人工知能学会全国大会での AI レクチャと研究会特別セッションで報告する．

1.5 ホームページの開設と運用

総括班としての活動の詳細を紹介するために，本特定領域のホームページ

<http://www.ar.sanken.osaka-u.ac.jp/activemining/>

を開設した．全体計画の概要，総括班会議，代表者会議，班会議の議題と議事録，研究会のプログラムなどの詳細を掲載している．また，各計画研究のホームページへのリンクが張ってあり，個別の研究の進捗状況，成果内容と成果も公開している．

1.6 アクティブマイニング特集号の企画

人工知能学会誌にアクティブマイニングの特集号を企画した．平成 14 年 9 月号に掲載予定である．本特定領域研究の概要と医療データ解析に関する解説記事の他に，本特定領域研究の成果を中心とする国内のアクティブマイニングの最新の研究成果が盛り込まれる予定である．

1.7 出版の企画

国内のアクティブマイニングの成果を世界に発信するため，1.4 で報告した研究会で発表された論文の出版化を企画した．全発表論文 42 件を査読し，27 件を選定し，IOS Press からアクティブマイニングに関する英文の編集本を出版する．出版は平成 14 年 9 月の予定である．また，本特定領域研究の計画研究代表者を中心にアクティブマイニングの教科書執筆を企画し，情報処理学会教科書委員会に企画書を提出した．

1.8 国際ワークショップの企画

来年度の活動を国際化するため，国際ワークショップを 2 件企画した．平成 14 年 8 月 18 〜 22 日に学術総合センターにて開催される第 7 回環太平洋人工知能国際会議に併設されるワークショップの一つにアクティブマイニングを主要課題とする知識獲得ワークショップを企画・提案した．また，平成 14 年 12 月 9 〜 12 日に前橋テルサにて開催される IEEE のデータマイニング国際会議に併設されるワークショップの一つにアクティブマイニングワークショップを企画・提案する．このワークショップに著名な海外研究者を招聘し，同時に本特定領域研究の外部評価を受ける．

1.9 研究成果発表会と報告書の刊行

平成 14 年 3 月 5，6 日に筑波大学ビジネス科学研究科大講義室にて本特定領域研究の全研究者の参加による研究成果発表の公開シンポジウムを開催した．発表会では総括班からの研究活動報告に続いて，各班の班長から班全体の研究活動の概要を報告し，ついで各計画研究代表者から個別の研究成果を報告し，最後に統括討論を行なった．研究成果発表会における予稿集もかねて，平成 13 年度研究成果報告書を刊行した．

2　各班の研究成果概要と来年度の計画

2.1　（A01 班）アクティブ情報収集

研究項目 A01 では，必要な情報リソースをシステム自らが能動的に探しに行く仕組みを確立することを目標とし，ネットワーク上に分散したデータベース内から使用者の目的や興味に関連のありそうな情報を，積極的に探索する技術やテキスト情報処理技術，それを後段のマイニング処理に引き渡す前処理を行うことを目的としている．特に，WWW におけるメタ情報源の獲得，分散動的情報からのアクティブ情報収集，多段階学習方式によるデータ収集と前処理の自動化，という三つの課題に焦点をあてて，研究を進めている．以下，各研究課題に関して，平成 13 年度の研究成果概要と平成 14 年度の研究計画についてまとめる．

1. WWW におけるメタ情報源の獲得

本研究課題の目標は，WWW などの膨大な情報空間に存在するメタ情報源を（半）自動的に収集し，それを利用することにより，従来手法よりも質と量，そして効率ともに飛躍的に向上するような情報収集を実現することである．これまでの WWW における情報収集とは，“Web ページの収集” を意味していた．これに対しこの課題では，有用な Web ページの所在や質に関する情報源である**メタ情報源**の収集を最終的な研究目的とする．メタ情報源には，情報の所在についての情報を提供する「LI 情報源」とコンテンツに関するメタデータを提供する「MD 情報源」がある．

平成 13 年度の研究成果

以下，具体的な研究テーマ毎に研究成果をまとめる．

1) **関係学習による対話的文書検索**

LI 情報源の収集のためには，インデックスページなどの特定が重要である．LI 情報源は，ハイパーリンクからなるグラフ構造をもち，またその Web ページのコンテンツにも特徴がある．そこで，**リンク構造**と**コンテンツ**からなる制約により，LI 情報源を判定できる．ある Web ページがある LI 情報源に属するか否かの判定は，クエリに対して文書が適合するか否かの判定と等価である．そこで，情報検索で広く使われている手法である適合フィードバックにより LI 情報源判定のための制約を学習させる実験を行い，良好な結果を得た．

2) **WWW 上の情報収集／可視化のための免疫ネットワークを用いたクラスタリング**

MD 情報源を獲得するには，あるユーザやユーザグループにおいて閲覧されている Web ページの系列に代表される**情報ストリーム**において，情報の流れの抽出，また情報の流れ間の関係を捉えることが必要になる．このような機能実現の要素技術として，免疫ネットワークを用いて文書集合間の時系列的関連を考慮したクラスタリング手法である**可塑的クラスタリング**を提案した．

3) **Web ページの部分更新のモニタリング**

収集された WWW における情報は，時々刻々変化する動的なものであるため，一旦獲得された情報源で頻繁に行われる更新を自動的に検知するシステムが必要不可欠になる．そこで，Web ページ上のユーザによって指定された一部分の情報に着目し，その部分に特定の更新があった場合のみ，その更新をユーザに提示する部分更新モニタリングシステムを開発した．このシステムでは，ユーザが監視させたい部分を特定するためのルールを帰納学習によって獲得する．また，特定された部分の更新がユーザにとって必要なものであるか判定するためのルールも学習する．これにより，メタ情報源などの獲得された情報源の更新を自動的に検知，通知することが可能になった．

平成 14 年度の研究目標

平成 14 年度は，これらの成果に基づき，1) LI 情報源の獲得のためのリンク構造制約の学習 2) 対話的文書検索システムへの SVM の適用に取り組む予定である．

2. 分散動的情報からのアクティブ情報収集

頻繁に更新される情報源からの情報収集を考えた場合，一時に更新されるそれぞれの情報量はそれほど大きくはない．我々はこのように少量で速報性のある情報オブジェクトを Ticker と呼び，それらを Web 情報源から収集し，統合することで利用者の意思決定や問題解決を支援する Intelligent Ticker システムを開発することを目標とする．

平成 13 年度の研究成果

Intelligent Ticker における情報抽出部のプロトタイプを開発した．プロトタイプでは二つの Web ページ入力に対して HtmlDiff を用いてその差分を抽出し，その差分の中で特に意味をもつ構造だけを抽出して表示するようにしている．抽出すべき有効な構造を選択するためには Web ページを木構造として解析し，変化のあった部分とその上位構造を抽出するようにしている．

これに基づき，航空便空席照会システムを開発した．その特徴は，以下のとおりである．

1) インターネット上に存在する国内航空三社のホームページから航空便の空席情報を収集する．この実現にはクエリ（搭乗日，出発空港，到着空港）の送出と結果ページから空席に関する情報抽出を行うラッパーを Java により記述している．

2) 空席照会は出発地から到着地への直行便だけでなく，収集した情報を統合することにより，乗り継ぎ便に関する情報も提供する．またこの乗り継ぎは異なった航空会社間の乗り継ぎも扱っている．

3) 情報収集には静的な情報を利用している．ここでの静的な情報は航空機の乗り継ぎ経路である．乗り継ぎ便も含めるとシステムは何度も情報検索を行う必要があるが，利用者が入力する希望到着時刻にできるだけ近く，また飛行時間が短い便の優先順位を高くして情報収集を行うようにしている．

平成 14 年度の研究計画

平成 14 年度においては，上記の成果を基礎にして，情報収集プランニング，静的な情報の更新，EBM への応用などを総合的に進める予定である．具体的には，構造化された情報を提供する医学分野として，医薬品の副作用情報の提供への応用を検討する．現在，医薬品情報に関しては厚生労働省を中心にその整備が行われようとしている．医薬品情報は副作用など生命にかかわる重要なものであるとともに，次々と新薬が開発される現在では頻繁に更新される情報の一つである．また治療に対して複数の医薬品が用いられることは一般的であり，その組み合わせによる副作用も無視できない．このように動的な情報源から複雑な要求を満たす情報検索を支援するために，アクティブ情報収集システムの応用は有効であると期待される．

3. 多段階学習方式によるデータ収集と前処理の自動化

従来からあるデータ処理技術では，分散した数多くのデータベースの中からあらかじめ必要なデータを選んで収集する作業を行う必要があった．これらの作業は「前処理」と呼ばれ，膨大な人手と時間を要していた．またデータに基づいて分類を行うだけで，対象間の関係を見出す能力も不足していた．以上の背景から，関連する情報を収集し，それらの関係を自動的に発見する手段として，データベース間の通信ネットワークに学習能力を持たせることを提案する．本研究の目的は，トポロジや情報伝達の優先順位を学習により動的に変化できる Global Intelligence Associating Network (GIANT) の構成法を明らかにした上で，現実のデータに基づいて動作するシステムを構築することにある．

平成 13 年度の研究成果

以下，具体的な研究テーマ毎に研究成果をまとめる．

1) **前処理支援システムの構築**

蓄積されたデータに対して前処理を施し，解析アルゴリズムが直接扱えるデータに変換する過程を明確に定義し，その過程に含まれる処理を効率的に行えるデータ構造，及び自動化アルゴリズムの提案と，実際のシステム構築を行った．具体的には，前処理で扱うデータをすべて XML 形式に統一し，利用者が処理の過程を記述する部分には XML から一意に変換されるデータ構造を用いて効率化を図り，利用者が一度作成したデータ変形フィルタについては自動的に並べ替えて利用者に提案する．

2) **伝言ゲーム型の情報収集**

情報収集および前処理は，情報提供者，ドメイン専門家，マイニング専門家の共同作業になることがほとんどで，それらの間で大量のデータが交換され，更新が頻繁に行われることに特徴がある．このような共同作業を支援するため，通常，メールと Web ページが用いられるが，不便な点が多い．データマイニングの作業はルーチンワークではないので，従来のグループウェアの適用も困難である．このような問題を解決する一つの方法として伝言ゲーム型の情報収集および前処理結果の交換を提案し，システムの実装と実験を行った．

3) **HTML のリンク構造と構文的特徴に基づく知識獲得**

現在の検索エンジンのテキスト照合の高速性を利用することで，構文的な特徴を共有する WWW ページを収集する．リンク集についてはバックリンクページがリンク集である頻度が高い事を利用し，データ集については予め XML の雛型を与えておくことによって検索エンジンを利用する．収集した WWW ページから知識を獲得するには，HTML 構文に関する制約を利用して知識を抽出する．これらの特徴を有するシステムを実装し，有効性を確かめた．

4) **高次元インデックス技術を用いた検索処理性能向上について**

ヒトの思考は空間的な直感に基づいていることが多い．その観点から，データマイニングの対象となるデータとして，位置情報を取り上げ，高次元インデックス技術を開発した．これにより，空間属性と非空間属性を併せもつ大量の時系列高次元データの前処理について，展望を得た．

平成 14 年度の研究計画

伝言ゲーム型の情報収集を実際に使ってもらい，有効性を評価すると共に改良すべき点を洗い出す．組み込まれている推薦機構について検討し，改良を行うと共に，前処理支援システムとの連携を図る．位置情報の処理を通じて，時系列高次元データの扱いについて研究を進める．A01 班の情報収集および前処理技術を集大成したシステムを設計する．

2.2 （A02 班）ユーザ指向アクティブマイニング

研究項目 A02 では，アクティブ情報収集機構で収集・前処理された情報から，ユーザの目的や興味に照らして重要・有用と思われる知識を発掘することを目標とし，多様な表現形式あるいは多種の情報源に対応できる「汎用性」，並びに，ユーザを含めた状況の変化に対応できる「柔軟性」の両者を兼ね備えた，大規模データからのマイニング手法の開発を目的としている．特に，構造化データからのマイニング，マイニングアプリケーション自動構築，例外性の発見，テキストデータからのマイニング，という 4 つの研究課題に焦点をあてて，研究を進めている．以下，各研究課題に関して，平成 13 年度の研究成果概要と平成 14 年度の研究計画についてまとめる．

1. 構造データからのアクティブマイニング

本研究課題の目標は，大規模なグラフ構造データ，空間分布構造データ，時系列構造データ，半構造データ，制約構造データなどの構造データを対象とし，ユーザの価値観を反映した重要なあるいは興味深い部分構造ならびにその特徴を知識として，ユーザの許容時間内に発掘するために必要な基礎技術を開発することである．以下，平成 13 年度の研究成果概要と平成 14 年度の研究計画についてまとめる．

平成 13 年度の研究成果

以下，具体的な研究テーマ毎に研究成果をまとめる．

1) **グラフ構造データからの特徴的な部分グラフの発見手法 GBI の性能向上**

逐次ペアのチャンキングというアイディアに基づき，構造の大きさに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改良した．分類性能に基づく指標を用いて部分構造を並行して抽出することにより，分類問題により柔軟に対応可能とした．発癌性化合物や変異源性化合物の分類に適用し良好な結果を得た．

2) **GBI 法の属性構築への適用**

決定木のノード生成時に分離能力最大のペアを選定し，チャンクして１つのグラフノードとし，各分岐ごとに同じ操作を再帰的に繰り返すことにより，分類に効果的な属性を逐次的に構築し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た．

3) **多頻度誘導部分グラフ抽出手法 AGM の連結部分グラフマイニングへの拡張**

実問題では連結部分グラフパターンを扱えれば解ける問題が多いことを考慮し，グラフ構造データからすべての多頻度誘導部分グラフパターンを高速導出する AGM 手法を，すべての多頻度連結部分グラフパターンを高速導出する AcGM 手法に拡張した．これにより，実規模の問題にも適用可能となった．

4) **文書クラスタリング**

対称性と反射律のみを許すラフ集合モデル TRSM を用いて文書を表現し，階層的クラスタリング，非階層的クラスタリングのアルゴリズムを開発した．それを情報検索に適用し良好な結果を得た．

5) **少数クラスのマイニング**

規則の帰納学習と相関規則マイニングを結合しクラス分布が非常に偏ったデータに対し，少数クラスを精度よく同定する手法を開発した．

6) **マイニング過程の視覚化**

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た．

7) **時系列データからのマイニング手法の開発**

経営データを対象に，従来の相関規則を拡張して，複数の時点における複数のブランド間の購買関連性の分析を可能にした．その結果，時系列データとして蓄積されている POS データの特徴を活かした，時間の経過にそった状態の遷移を表現することが可能になり，従来，十分に取り扱うことができなかった複雑な社会現象を時系列データとして分析することができる可能性が出てきた．

平成 14 年度の研究目標

平成 14 年度は，これらの成果に基づき，1) グラフ構造データマイニング手法の並列探索化と医療データへの適用，2) 属性構築型グラフ構造データ分類手法の精度評価と改良，3) 多頻度連結グラフマイニング手法の性能評価と改良，4) 密度関数法と類似性によるクラスタリング手法の開発とテキストマイニングへの適用，5) 少数クラスマイニング手法の医療データへの適用，6) 構造データの可視化，7) 種々のビジネス領域（食品，アパレルなど）における競争ブランド群の統一的な時系列マイニング手法の枠組み構築，に取り組む予定である．

2. メタ学習機構に基づくアクティブマイニング

本研究課題の目標は，マイニングシステムの開発プロセスを詳細に分析し，その中で特に重要な「データ加工」と「データマイニング」プロセス，並びに「ユーザの主観的基準」をリポジトリ化し，メタ学習機構に基づいて，これらのリポジトリを統合しながら，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ツールを開発することである．以下，平成 13 年度の研究成果概要と平成 14 年度の研究計画についてまとめる．

平成 13 年度の研究成果

以下，具体的な研究テーマ毎に研究成果をまとめる．

1) **リポジトリに基づく帰納アプリケーション構築支援環境**

従来，代表的な帰納アルゴリズムを分析し，帰納メソッドリポジトリを構築してきたが，コミッティ学習メソッドを追加するとともに，ポルト大学の StatLog プロジェクトから提供されている 8 種類のデータセットを利用して，帰納アプリケーションの自動合成実験を行った．その結果，本ツールで合成された帰納アプリケーションの平均正解率は，StatLog プロジェクトで調査された 24 種類の代表的な帰納アルゴリズムのどの平均正解率よりも高い値を示した．さらに，効率的な仕様探索を実現するために，相関ルールに基づいて，仕様書換ルール（メタルール）の学習を試みた．その結果，ランダム探索よりは安定した仕様書換が実現できることが確認されたが，最良の仕様を短時間に見つけ出す点については不十分であり，メタルール自身の構造について課題が残った．

2) **慢性肝炎データセットのデータ前処理とルール発見**

本研究領域で対象となっている慢性肝炎データセットを分析し，離散値に基づくルール発見を試みた．まず，出現頻度に基づいて，957 種類の検査項目を 41 種類に絞り込んだ．次に，検査周期については，患者に依存して異なっており，これも出現頻度に基づいて 28 日周期に統一し，空値になるデータ項目については，線形補完によりその値を推定した．また，Das の手法に基づき，時系列データをある一定長で切り出し，ＥＭアルゴリズムに基づき，切り出されたサブシーケンス群のクラスタリングを行った．以上の準備の下に，決定木学習により，検査項目値から GPT の 1 年間の変化を予見するルールの学習を試みた．その結果を専門医に評価してもらった所，ルールに

よっては理解できないといったコメントもあったが，乳びとＧＰＴの関連性，ＧＰＴの周期性の予測など，興味深いというコメントもいくつかもらえた．

3) **多粒度の問題解決メソッドの検討**

「慢性肝炎データセットのデータ前処理とルール発見」において，一次の生データからマイニングへの入力データへの変換過程には，いくつもの人手による作業が含まれており，その作業プロセスを自動化することも，アプリケーション開発現場では，大いに有益である．そのような解決に向けて，本テーマでは，要求レベル－問題解決レベル－実装レベルと対応づけた 3 層のメソッドリポジトリを検討し，ビジネスドメインにおいて，プロトタイプを試作し，その有用性を確認した．

4) **領域オントロジー構築支援ツール**

ユーザにとって重要・有用な知識を発見するには，マイニングだけでなく，モデリング的な機能も必要になってくるものと予想している．そのため，ドメインモデルを構成する概念要素の仕様を与えるドメインオントロジーを構築するツールの開発を手がけ，計算機可読型辞書とテキストコーパスから，ツールの試作を行った．

平成 14 年度の研究計画

平成 14 年度においては，上記の成果を基礎にして，マイニングリポジトリの拡張，属性選択の考察，データ前処理メソッドのリポジトリ化，ユーザの興味指標のリポジトリ化，医学用語のオントロジー化などを総合的に進める予定である．

3. 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

例外性は，大多数の傾向とは異なる性質を表し，それ自体が興味深いことに加えて新しい知識発見の糸口となることが多く，データマイニングにおいて有用知識の発見に特につながりやすい発見知識候補として重視されてきた．従来手法は種々の問題に対して成功を収めたが，例外性がデータと知識だけに関して定義されており，ユーザの解析目的や社会状況などの環境に関する側面が無視されている．本研究課題では，データ，知識，および環境などに関する例外性を連鎖的に発見するアクティブマイニング手法を構築し，手法に基づく実装システムを医学・商業・科学データなどに適用して領域専門家による評価で有効性を実証する．さらに例外性発見は，状況変化に応じて能動的に有用知識を発見するアクティブマイニングの必須技術であり，本研究課題でもこの機能を実現する．以下，平成 13 年度の研究成果概要と平成 14 年度の研究計画についてまとめる．

平成 13 年度の研究成果

以下，具体的な研究テーマ毎に研究成果をまとめる．

1) **スパイラル的例外性発見手法**

既に発見された例外性をデータ，知識，および環境に照らしあわせ，新しい例外知識を発見する手法を開発した．データ，知識，および環境はそれぞれ，与えられたデー

タ，既に発見された知識，および妥当性・有用性・新規性・意外性に関するユーザの評価値と定義される．発見手法は，既発見知識の近傍を，異なる離散化手法と確率的規準に基づいて探索する．提案手法はデータマイニングの標準問題として提供されている髄膜炎データに適用済みである．

2) **状況変化の検知手法**

統一的な規準を確立するために，ルール発見の新しい最悪解析を行った．この解析は従来研究と異なり，ルール発見に重要である複数の指標に関する誤差が指定値以内である場合を，指定した確率以上とするために必要な例数を与える．提案する PAGA (Probably Approximately General and Accurate) 発見は，分類学習において標準的な PAC (Probably Approximately Correct) 学習の自然な拡張ともなっている．現在 PAGA 発見に基づき，発見された例外性を用いて，データ，知識，および環境に関する状況変化を検知する手法を開発中である．

3) **スパイラル的特異ルール発見手法**

対象データを変更し，興味深い特異データを抽出することにより，特異ルールを発見する手法を開発した．抗原抗体反応に関するアミノ酸配列データなどに適用し，提案法の有用性を示した．また，特異ルールを相関ルールや例外ルールと形式的に比較・分析し，特異指向マイニングの理論的根拠を確立した．現在，マイニングプロセスのメタレベルの制御メカニズム，複数のマイニングエージェント，マルチデータソースからのマイニング手法を開発中である．また，実験用の複数分野のデータを収集・準備中である．

なお紙面の都合上説明は省くが，本課題の基盤技術として有用な機械学習とデータマイニングに関しても成果をあげた．

平成 14 年度の研究計画

今年度，残された時間を用いて状況変化の検知手法と，共通データとしてプロジェクトより提供された肝炎データの解析に取り組む．来年度は当初の計画通り，スパイラル的例外性発見におけるスケジューリング手法と分散協調型特異ルール発見手法に取り組む予定である．

4. 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

本研究課題では，ある特定の意味クラスに属する用語の発見を目的とし，論文要旨に出現する名詞句がそのクラスの用語であるかどうかを同定するタスクとして問題設定を簡略化し，学習に基づく手法を考案することが目標である．特に，学習事例が少ないという現実的な想定の基に，用語が出現するテキスト中の文脈が意味クラス推定にどの程度有用であるかを明らかにすることを目指した．以下，平成 13 年度の研究成果概要と平成 14 年度の研究計画についてまとめる．

平成 13 年度の研究成果

以下，具体的な研究テーマ毎に研究成果をまとめる．なお，これらの研究は，蓄積された正しいデータからの学習に基づくシステムの構築を目指し，学習手法として Support Vector Machines を利用した．

1) **未知語を含む英文文書内の単語の品詞推定**

専門分野の文書には一般の辞書には含まれない語が多く出現し，品詞等の文法情報の特定に支障をきたす．専門性の高い分野には次々に新しい用語が出現するため，それらをすべて辞書に登録することは現実的に不可能である．そこで，前後の文脈，あるいは，単語の綴り (特に接頭，接尾表現) を手がかりとして，未知の単語の品詞を決定し，それによって，専門用語と考えられる名詞句の同定を柔軟に行うことを試み，従来の手法より高い解析性能を示すことができた．

なお，本システムで利用した SVM は，高い学習能力を有するものの，学習および実行に時間がかかり，実用面では問題があった．そのため，第一の処理として，従来型の N-gram に基づく統計的品詞付与システムによる学習を行い，そのシステムが誤りを生じる箇所を SVM によって学習するという二段構えの方法を提案した．これにより，解析精度を維持しながら，実用的な解析時間で未知語を含む専門文書中の単語の品詞推定を約 97%という高い精度で達成することができた．

2) **文書中の基本句の自動抽出**

品詞付与が行われた文書に対して，そこに現れる名詞句や動詞句などの基本句を精度よく同定することは，その後の言語処理の性能に大きく影響を与える．特に名詞句の同定は，専門用語の抽出にとって必須の処理である．英語の基本句の同定を各単語へのラベル付与問題として見直し，SVM を用いた学習システムを複数混合することにより，従来法を上回る精度での基本句抽出を達成した．

3) **テキストからの専門用語の抽出と分類**

Medline アブストラクトを題材とし，テキスト中に現れる名詞句を対象にして，その意味クラスを推定するタスクに関する実験を行った．ある程度教師なしの手法を実践することを目指し，コーパスの精度は落ちるものの，次のように人手による作業をなるべく少なくしつつ，より大規模なデータに対して，特定の意味クラスの語の推定実験を行った．

今回の実験では，病名のみに注目し，未知の名詞句に対し，それが病名であるか否か自動推定可能かどうかを確認することを目的とした．用語のバリエーションがある程度限定されてしまっていると考えられるので，用語の内部情報を過度に利用するのは避けることにし，文脈情報によってどの程度用語の意味クラス推定が可能かを確認することを主たる目的とした．
その結果，文脈情報のみを用いた場合は，精度が約 70%，再現率が約 60%で病名かど

うかを判定することができた．また，用語内の文字情報まで利用する場合には，精度が約 90%，再現率が約 80%で判定が可能であることがわかった．

平成 14 年度の研究計画

今回用いた未知語処理を伴う品詞推定，および，基本句へのまとめ上げプログラムは，現在発表されているシステムの中では最も高い精度を示しており，現時点では充分優れたものである．しかし，それぞれのシステムは，現在入手可能なタグ付きコーパスである Penn Treebank から学習を行ったものであり，今回利用した医学生物学分野の論文とは，内容が大きく異なる．今後，類似分野のタグ付きコーパスを蓄積することによって，より精度の高い解析を行える可能性がある．また，今回は利用できなかったが，英語の単語あるいは句単位の係り受けに対しても学習手法の適用を初めており，今後，このような高精度な言語処理システムを用いた実験を行うことが課題である．今回の実験では，病名のみの判定を課題としたが，薬品名や治療法など，医学文献を解析する上で重要なクラスの用語がある．クラスの分類をより詳細化した場合への本手法の適用についても今後の課題である．

2.3 （A03 班）アクティブユーザリアクション

研究項目 A03 では，具体的な問題領域（医療，化学薬品）を対象にデータマイニングシステムを構築し，アクティブマイニングの結果得られた知識を，適用領域のユーザにとって有用なものとするための仕組を具体化することを目標としている．その実現のために，発掘された知識の表示法，評価手法 (有用性，新規性，意外性など)，ユーザからの効果的なフィードバックの手法など，ユーザのアクティブなリアクションを容易にし，“科学的発見のらせんモデル” を実現するユーザのアクティブな新データ収集，設定目的変更などを容易にする一般的な枠組みの構築に取り組んでいる．具体的には，診療情報生成システムの開発，化学物質群からのリスク分子発見およびヒューマン・システム・インタラクションに基づく知識評価と選択という 3 つの研究課題に焦点を当て，研究を進めている．以下，各研究課題に関して，平成 13 年度の研究成果概要と平成 14 年度の研究計画について述べる．

1. 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

本研究課題は，ラフ集合に基づくあいまいな知識の取り扱い方法を利用，専門医のあいまいな知識を定量的に取り扱えるようにして，専門家の知識との総合作用から知識の発掘の促進および医療現場に有用な診療情報の生成を目標としている．特に，本年度は共通データが時系列データであることに着目し，時系列データからの診療情報生成に関する基礎的技術の開発を行った．

平成 13 年度の研究成果

1) **多重スケールマッチングによる時系列データのセグメント化**

病院情報システムに蓄積された膨大な時系列データでは，同一患者の検査データを数年から数十年の長期にわたり継続的に収集した時系列検査データが利用可能になりつつある．このような時系列検査データは，数日を単位とする短期間の推移のみならず，

年単位の長期にわたる検査値推移パターンと疾患との対応関係を示すものであるため，その解析により慢性疾患を誘発する要因の特定，あるいは発病時期の予測等が可能になると期待される．しかしながら，これらのデータは当初から解析を目的に収集されたものではなく，不均質なものであるため，解析が困難である．本研究では，データの平滑化とともにセグメント化し，セグメント間でのマッチングを容易にする方法として多重スケールマッチングを用いた．

2) **ラフクラスタリングによる系列データの分類**

対象間の類似度が原点を持たない相対的類似度として与えられる場合，クラスタ内の分散，重心等を定義することが困難で，クラスタとしてのまとまりを評価することは容易ではない．ラフクラスタリングは，ラフ集合論の識別不能性の概念に基づくクラスタリング法であり，対象のまとまり具合いを識別不能度として表現することで，相対的類似度で表現されたデータにおいても可読性の高いクラスタを生成することができる．本研究では，ラフクラスタリングを用いて，多重スケールマッチングでセグメント化した系列を分類する方法を，共通データを含めた病院情報システムのデータの一部に適用し，良好な結果が得られた．

平成 14 年度の研究目標

平成 14 年度は，これらの成果に基づき，1) 上記手法の共通データへの適用，2) 他の医療データへの適用と精度評価，3) 多重スケールマッチングおよびラフクラスタリングの性能評価と改良を行いつつ，時系列データにおけるユーザリアクションを実現するための新たな手法の開発を行っていく予定である．

2. 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

本研究課題は，薬品類とその活性のデータベースから，1) それぞれの生理活性に対して特徴的な部分構造や物理化学的性質を見出し，2) 新規の化学物質群から未知の生理活性を予測して，危険を回避することを目的としている．本年度の研究では，変異原性と発ガン性に関しては，Apriori-based graph mining およびカスケードモデルを，ドーパミン活性に関しては，分子グラフの類似性に基づく事例ベース推論を適用した．

平成 13 年度の研究成果

1) **変異原性に対する検討**

芳香族ニトロ化合物の変異原性問題に Apriori-based graph mining およびカスケードモデルを適用した．その結果から,「芳香族ニトロ化合物で ortho 位置換基の有無が重要な因子となる」というような有意義な結果が得られた．

2) **発がん性の解析と予測**

発がん性の解析と予測には，カスケードモデルを適用した．この適用結果から,「有機塩素化合物の活性と水素結合受容体の有無および分子の柔軟性との間の高い相関」を

見出すことができた．なお，発ガン性の解析と予測は，国際ワークショップ Predictive Toxicology Challenge 2001 に参加して発表したが，モデルの理解容易性において参加 14 グループ中第１位の評価を受け，また female rat に対する予測精度においても，ROC 解析の結果が非常に良いとの評価を受けることができた．

3) **ドーパミン活性に関する検討**

各種の性質にもとづくトポロジカルフラグメントスペクトルを利用して，化合物構造間の距離を定義する．どのような性質が活性の発現とよく対応するかの検討を進めた．適用結果から，分子グラフの類似性にもとづく事例ベース推論では，化学者が見て納得できる類似構造の分子を選択することができた．ドーパミン活性については，類似 20 分子中で 3 分子が活性を示した．

平成 14 年度の研究目標

平成 14 年度は，これらの成果に基づき，1) 既存薬品の全て（およそ１２万種）を網羅するデータベース MDDR と毒性のデータベース NCI（およそ２１万種）を使用，対象生理活性に関する解析，2) カスケードモデルのルールの理解容易性の向上，3) 粒度の細かい知識発見のための Apriori based graph mining の改良，4) 事例ベース推論における類似性規範の洗練に取り組む予定である．

3. 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

「知識が有用である」とは，利用者にとって理解が容易であり目的に応じて的確に使用できること，また，利用者の創造性を刺激しうる機能を備えることを意味する．そのために，本研究課題では，知識の需給関係に注目し，知識の候補を供給するシステムとそれを解釈・選択・利用する専門家とのインタラクションを通じて，知識を評価・選択できるような方式を確立することが目標であり，この方式は，利用者個人あるいはグループの主観までを評価尺度に含み，従来研究されてきた客観的な基準での知識評価方法を超えるものとする．

平成 13 年度の研究成果

以下，具体的な研究テーマ毎に研究成果をまとめる．

1) **客観的意志決定の支援: 病態モデルの構築**

血液中の複数の酵素データを用いて病態モデルを臨床検査の実データから因子分析を用いて構築し，因子スコアの視覚的表現を試みた．視覚化によって，臨床家の検査データ評価に関わる負担を軽減し，さらに，重要な病的異常値の見逃しを防止する効果が得られた．

2) **客観的意志決定の支援: 病態パターンのクラスタリング**

血清蛋白分画データの病態パターンのクラスタリングを自己組織化マップを用いて行い，得られたクラスターを病態と対応付けるために，分類クラスと他の臨床検査デー

タを属性として決定木分析を行った．意味付けされた分類クラスには，臨床診断につながる付加価値情報として有用な情報が得られた．

3) **主観的発見: チャンス発見プロセスの実現**

チャンスを評価するメタな指標として，P: 行動の提案可能性，U: 気付かれにくさ，G: 成長性を提案した．P と G を具体的に把握するために，一人ではなく数人のグループを構成し，グループディスカッション（GI）における人と人の間でのチャンスに基づく提案や採択を通してチャンスを選択してゆくプロセスを実現した．さらにこのプロセスを，意思決定環境についてのデータ（環境データ：販売者における購買データなど）に適用し，視覚的データマイニングによって刺激してチャンスを発見する仮定を促進する結果を得た．

4) **主観的発見: チャンス発見二重らせんモデルの実現**

新たなチャンスを発見するモデルとして，二重らせんモデルを提案し，これを実現するためのヒューマン・マシンインターアクションを促進するシステムを構築した．具体的には，グループインタビュー（GI）における議論の各参加者の考えを，カードに思い当たるだけ書きこんでもらい，そのテキスト内の単語の相関関係を図示する視覚的なテキストマイニング（KeyGraph）の出力図によって議論におけるチャンス表出化を促進するシステムを構築した．このシステムで顧客像を把握した結果，顧客の購買（POS）データからスーパーマーケットの購入金額増加の鍵となる商品や，その店の経営状態のおおまかな変化を示す予兆を発見することができた．さらに，社会調査を行った社会学者の「関心」を起点として二重螺旋モデルを実行した結果，社会学的に新規性と説明能力の高い仮説が獲得された．

平成 14 年度の研究目標

平成 14 年度は，これらの成果に基づき，さらに積極的に人やそのグループの判断をシステムに取り入れ，それにより発見される外界の未知要因の重要さを計る指標とその発見に至るプロセスの支援方法を確立するための研究を進めていく予定である．

3 成果リスト

今年度発表の成果リストを示す．

3.1 ジャーナル論文

[1] 岡部正幸，山田誠二：関係学習を用いた対話的文書検索，人工知能学会誌，Vol. 16, No. 1, F (2001).

[2] 山田誠二，村瀬文彦：ページ情報エージェントの組織化による Web ブラウザの適応インタフェース，人工知能学会誌，Vol. 16, No. 1, F (2001).

[3] 北村泰彦，野田知哉，辰巳昭治：動的情報メディエータのための知的情報収集手法，電子情報通信学会論文誌 D-I, Vol. J84-D-I, No. 8, pp. 1256-1265 (2001).

[4] Chowdhury Rahman Mofizur and Masayuki Numao: Automated bias shift in a constrained space for logic program synthesis, 人工知能学会論文誌, Vol. 16, No. 6, pp. 548-556 (2001).

[5] T. Matsuda, H. Motoda, T. Washio: Graph-Based Induction and Its Applications, to appear in *AI in Engineering* (2002).

[6] 松田　喬, 元田　浩, 鷲尾　隆: 一般グラフ構造データに対するGraph-BasedInductionとその応用，　人工知能学会誌，Vol. 16, No. 4, pp. 363-374 (2001).

[7] 寺邊　正大, 鷲尾　隆, 元田　浩: S^3Baggingによる高速な分類器生成, 数理モデル化と応用，Vol. 42, No. 14, pp. 25-38 (2001).

[8] 堀 聡，瀧 寛，鷲尾 ，元田 浩: データマイニングを用いた市場品質監視システム, 電気学会 電子・情報・システム部門誌, Vol. 121-c, No. 8 (2001).

[9] T.B. Ho, T.D. Nguyen, D.D. Nguyen, and S. Kawasaki: Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining, *International Journal of Artificial Intelligence Tools*, pp. 691-713 (2001).

[10] 矢田勝俊, 羽室行信, 加藤直樹: 経営データからの知識発見, 国民経済雑誌, Vol. 184, No. 1, pp. 19-33 (2001).

[11] Y. Hamuro, E. Kawata, N. Katoh and K. Yada: A Machine Learning Algorithm for Analyzing String Patterns Helps to Discover Simple and Interpretable Business Rules from Purchase History, to appear in *Progresses in Discovery Science, State-of-the-Art Surveys*, LNCS, Springer-Verlag.

[12] 和泉憲明, 山口高平: オントロジーに基づくソフトウェアエージェントのパターン指向開発, 電子情報通信学会誌, Vol. J-84-D-I, No. 8, pp. 1181-1190 (2001).

[13] 和泉憲明, 山口高平: 教師支援エージェント構築のためのタスクパターンリポジトリの開発, 教育システム情報学会誌, Vol. 18, No. 3・4, pp. 352-363 (2001).

[14] H.Abe and T.Yamaguchi: Constructing Inductive Applications byMeta-Learning with Method Repositories, to appear in *Progresses in DiscoveryScience, State-of-the-Art Surveys, LNCS*, Springer-Verlag.

[15] N. Zhong, C. Liu, and S. Ohsuga: Dynamically Organizing KDD Process, *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 15, No. 3, World Scientific, pp. 451-473 (2001).

[16] N. Zhong, J.Z. Dong, and S. Ohsuga: Rule Discovery by Soft Induction Techniques, *Neurocomputing, An International Journal*, Vol. 36 (1-4) Elsevier, pp. 171-204 (2001).

[17] N. Zhong, J.Z. Dong, C. Liu, and S. Ohsuga: A Hybrid Model for Rule Discovery in Data, *Knowledge Based Systems, An International Journal*, Vol. 14, No. 7, Elsevier, pp. 397-412 (2001).

[18] N. Zhong, J.Z. Dong, and S. Ohsuga: Using Rough Sets with Heuristics to Feature Selection, *Journal of Intelligent Information Systems*, Vol. 16, No. 3, Kluwer, pp. 199-214 (2001).

[19] N. Zhong and A. Skowron: A Rough Sets Based Knowledge Discovery Process, *International Journal of Applied Mathematics and Computer Science*, Vol. 11, No. 3, Technical University Press, pp. 101-117 (2001).

[20] N. Zhong: Rough Sets in Knowledge Discovery and Data Mining, *Journal of Japan Society for Fuzzy Theory and Systems*, Vol. 13, No. 6, pp. 581-591 (2001).

[21] C. Liu and N. Zhong: Rough Problem Settings for ILP Dealing with Imperfect Data, *Computational Intelligence, An International Journal*, Vol. 17, No. 3, Blackwell Publishers, pp. 446-459 (2001).

[22] S. Yokoyama, K. Matsuoka, S. Tsumoto, M. Harao, T. Yamakawa, K. Sugahara, C. Nakahama, S. Ichiyama, and K. Watanabe: Study on the Association between the Patient's Clinical Background and the Anaerobes by Data Mining in Infectious Disease Database, *BMFSA*, Vol. 7, No. 1, pp. 121-128 (2001).

[23] 安田 晃, 柳樂真佐実, 孫 暁光, 津本周作, 山本和子: 自主学習における学生の自己評価の変動に関する解析, 医学教育, Vol. 32, pp. 69-75 (2001).

[24] 津本周作: 医学における知識発見手法の可能性 (特集: データマイニングコンテスト), 情報処理, Vol. 42, pp. 472-477 (2001).

[25] 津本周作: ラフ集合論の現状と課題 (特集: ラフ集合の理論と応用), 日本ファジィ学会誌, Vol. 13, pp. 552-561 (2001).

[26] 鈴木英之進, 菅谷信介, 津本周作: サポートベクターマシンに基づく医療データからの事例発見, オペレーションズ・リサーチ, Vol. 46, No. 5, pp. 243-248 (2001).

[27] T. Okada: Discovery of Structure Activity Relationships using the Cascade Model: The Mutagenicity of Aromatic Nitro Compounds, *Journal of Computer Aided Chemistry*, Vol. 2, pp. 79-86 (2001).

[28] A. Inokuchi, T. Washio, T. Okada, and H. Motoda: Applying the Apriori-based Graph Mining Method to Mutagenesis Data Analysis, *Journal of Computer Aided Chemistry*, Vol. 2, pp. 87-92 (2001).

[29] 大澤幸生, 高間康史: Web に潜む創造的意思決定のチャンス, 人工知能学会誌 Vol. 16, No. 4 pp. 530-534 (2001).

[30] Y. Ohsawa: Chance Discoveries for Making Decisions in Complex Real World, *New Generation Computing* (Springer Verlag and Ohmsha), Vol. 20 No. 2 (2002).

[31] W. Sunayama, Y. Nomura, Y. Ohsawa and M. Yachida: Support System for User Interests Expression on Searching Web Page, *Systems and Computer in Japan* Vol. 32, No. 13, pp. 14-22 (2001).

[32] 松村真宏, 大澤幸生, 石塚満: テキストによるコミュニケーションにおける影響の普及モデル, 人工知能学会論文誌 Vol. 17 (印刷中) (2002).

[33] 高田真好，寺野隆雄：2 段階 CLP 緩和解法によるリソース平準化スケジューリングシステム．電子情報通信学会論文誌 D-I, Vol. J84-D-I No. 6, pp. 896-905 (2001).

[34] 倉橋節也，寺野隆雄：エージェントシュミレーションによる共同分配規範モデル．電子情報通信学会論文誌 D-I, Vol. J84-D-I No. 8, pp. 1160-1168 (2001).

[35] K. Takadama, T. Terano, K. Shimohara: Non-Governance Rather Than Governance in a Multiagent Economic Siciety, *IEEE Transaction on Evolutionary Computing*, Vol. 5, No. 5, pp. 535-545 (2001).

[36] R. Kudo, T. Terano: Automation of Concept Development, *Expert Systems*, Vol. 6 No. 45, pp. 1641-1665 (2002).

3.2 国際発表

[1] M. Okabe and S. Yamada: Interactive Web Page Filtering with Relational Learning, *The First Asia-Pacific Conference on Web Intelligence* (WI-2001), pp. 443-447 (2001).

[2] S. Yamada and Y. Osawa: Information Gathering of Web pagesto Guide Concept Understanding, *The Tenth International World Wide Web Conferenc* (WWW10), Poster 1046 (2001).

[3] N. Nagino and S. Yamada: Assistance of Web browsing by indicating the future Web pages, *The 9th International Conference on Human-Computer Interaction*, pp. 1140-1144 (2001).

[4] M. Mase and S. Yamada: Development and Evaluation of an Information Retrieval System for User Groups and The WWW, *The 9th International Conference on Human-Computer Interaction*, pp. 1125-1129 (2001).

[5] M. Okabe and S. Yamada: Interactive Document Retrieval with Relational Learning, *SAC-2001 ACM SYMPOSIUM ON APPLIED COMPUTING*, pp. 27-31 (2001).

[6] Y. Takama and K. Hirota: Consideration of Presentation Timing Problem for Chance Discovery, *5th World Multiconference on Systems, Cybernetics and Informatics* (SCI2001), pp. 429-432 (2001).

[7] Y. Takama, M. Kawabe, K. Hirota: Kansei-keyword Extraction from Japanese Film Scenario Using Sensitivity Information, *Joint 9th IFSA World Congress and 20th NAFIPS International Conference* (IFSA/NAFIPS2001), 513(CD-ROM Proceedings) (2001).

[8] Y. Takama and K. Hirota: Employing Immune Network Model for Clustering with Plastic Structure, *2001 IEEE International Symposium on Computational Intelligence in Robotics and Automation* (CIRA2001), MP-6-1 (CD-ROM Proceedings) (2001).

[9] Z. Stejic, Y. Takama, K. Hirota: Integrated Querying of Images and Text, *2nd Int. Symposium on Advanced Intelligent System Conference* (ISIS2001), pp. 66-70 (2001).

[10] E. M. Iyoda, Z. Stejic, Y. Takama, K. Hirota: Image Retrieval Using Local Similarity Patterns Inferred by Genetic Algorithm, *2nd Int. Symposium on Advanced Intelligent System Conference* (ISIS2001), pp. 101-105 (2001).

[11] Y. Takama, K. Oh, K. Hirota: Finding landmarks in Keyword Map Based on Immune Network, *2nd Int. Symposium on Advanced Intelligent System Conference* (ISIS2001), pp. 37-41 (2001).

[12] Z. Stejic, Y. Takama, K. Hirota: Improving Retrieval Effectiveness by Integrated Querying of Images and Text, *12th Int'l Conf. On Information and Intelligent Systems* (IIS2001), pp. 233-244 (2001).

[13] Y. Takama and K. Hirota: Finding Topic Distribution from A Sequence of Document Sets, *2nd Vietnam-Japan Symposium on Fuzzy Systems and Applications* (VJ-FUZZY), pp. 132-139 (2001).

[14] Y. Takama and K. Hirota: WWW Information Visualization Based on Plastic Clustering, *10th IEEE International Conference on Fuzzy Systems* (FUZZ-IEEE2001), S404 (CD-ROM Proceedings) (2001).

[15] Z. Stejic, E. M. Iyoda, Y. Takama, K. Hirota: Automatic Textual Summarization of Image Database Contents Using Combination of Clustering and Neural Network Techniques, *2nd International Conference on Intelligent Technologies* (InTech'2001), pp. 233-239 (2001).

[16] Y. Takama and K. Hirota: Consideration of Memory Cell for Immune Network-based Plastic Clustering Method, *2nd International Conference on Intelligent Technologies* (InTech'2001), pp. 409-414 (2001).

[17] Yasuhiko Kitamura, Teruhiro Yamada, Takashi Kokubo, Yasuhiro Mawarimichi, Taizo Yamamoto, Toru Ishida: Interactive Integration of Information Agents on the Web (Invited Talk), Matthias Klusch, Franco Zambonelli (Eds.), Cooperative Information Agents V, Lecture Notes in Artificial Intelligence 2182, Berlin et al.: Springer-Verlag, pp. 1-13 (2001).

[18] Satoshi Oyama, Takashi Kokubo, Teruhiro Yamada, Yasuhiko Kitamura, Toru Ishida: Keyword Spices: A New Method for Building Domain-Specific Web Search Engines, *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (IJCAI-01), pp. 1457-1463 (2001).

[19] W.E. Walsh, M. Yokoo, K. Hirayama, M. P. Wellman: On Market-Inspired Approaches to Propositional Satisfiability, *Proceeding of the 17th International Joint Conference on Artificial Intelligence* (IJCAI-01), pp. 1152-1158 (2001).

[20] Miyako Tanaka, Sanae Nakazono, Hiroshi Matsuno, Hideki Tsujimoto, Yasuhiko Kitamura, Satoru Miyano: Intelligent System for MEDLINE Record Selection by Keyword Recommendation and Learning Text Characteristics, *Pacific Symposium on Biocomputing 2001*, pp. 130 (2001).

[21] Kawano, H.: Architecture of Trip Database Systems: Spatial Index and Route Estimation Algorithm, *XIV International Conf. of Systems Science*, Vol. III, pp. 110-117, Poland (2001).

[22] Cholwich Nattee and Masayuki Numao: Geometric method for document understanding and classification using on-line machine learning, *Sixth International Conference on Document Analysis and Recognition*, pp. 602-606. IEEE Computer Society Press (2001).

[23] Tuan Nam Tran and Masayuki Numao: Text data mining in biomedical literature by combining with an information retrieval approach, *Proc. the 14th International Conference on Applications of Prolog*, pp. 295-304. INAP Organizing Committee/Prolog Association of Japan, (2001).

[24] Masayuki Numao, Daishi Kato, and Masaru Yokoyama: Learning organization in global intelligence. *AAAI Spring Symposia*, AAAI Press (2002).

[25] T. Wada, H. Motoda and T. Washio: Knowledge Acquisition from Both Human Expert and Data, *Proc. of the Fifth Pacific-Asia Conference on Knowledge Discovery and Data Mining* (PAKDD01), Lecture Notes in Artificial Intelligence, Springer-Verlag, pp. 550-561 (2001).

[26] T. Washio and H. Motoda: Discovery of Law Equations governing Human Affinity under Trade-off between Cost and Risk *Proc. of International Meeting of The Psychometric Society* (IMPS-2001), pp. .74 (2001).

[27] M. Terabe, T. Washio and H. Motoda: S^3Bagging: Fast ClassifierInduction Method with Subsampling and Bagging, *Advances in Intelligent Data Analysis, Proc. of the Fourth International Symposium on Intelligent Data Analysis*, Springer, pp. 177-186 (2001).

[28] M. Terabe, T. Washio and H. Motoda: The Effect of Subsampling Rate on S3 Bagging Performance, *Proc. of Active Learning, Database Sampling, Experimental Design: Views on Instance Selection*, Workshop of ECML/PKDD2001, pp. 48-55 (2001).

[29] T. Washio and H. Motoda: Discovering Admissible Simultaneous Equation Models from Observed Data, *Machine Learning: ECML2001, Proc. of the 12th European Conference on Machine Learning*, Springer, pp. 539-551 (2001).

[30] T.D.Nguyen, T.B. Ho, and H. Shimodaira: A Scalable Algorithm for Rule Post-Pruning of Large Decision Trees, *5th Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD00*, Hongkong. Lecture Notes in Artificial Intelligence 2035, Springer, pp. 467-476 (2001).

[31] T.B. Ho, S. Kawasaki, and N.B. Nguyen: Cluster-based Information Retrieval with Tolerance Rough Set Model, *Second International Symposium on Advanced Intelligent Systems*, Daejeon, Korea, pp. 6-11 (2001).

[32] T.B. Ho, D.D. Nguyen, and S. Kawasaki: Mining Prediction Rules from Minority Classes, *14th International Conference on Applications of Prolog* (INAP2001), International Workshop Rule-Based Data Mining RBDM pp. 254-264 (2001).

[33] T.B. Ho, T.T. Nguyen, P.C. Nguyen, and C.M. Luong: Towards a Practical Framework for Vietnamese Natural Language Processing, *Vietnam-Japan Symposium on Fuzzy Systems and Applications*, pp. 297-304 (2001).

[34] Y. Hamuro, N. Katoh and K. Yada: Discovering Association Strength among Brand Loyalties from Purchase History, *Proceeding of 2001 IEEE International Symp. on Industrial Electronics*, pp. 114-117 (2001).

[35] Y. Hamuro, N. Katoh, E.H. Ip, S.L. Cheung, K. Yada: Discovery of Interesting Rules for Purchase Behavior Using String Pattern Analysis, *The 2001 World Congress of Mass Customization and Personalization* (MCPC 2001) (2001).

[36] Y. Hamuro, N. Katoh, K. Yada and T. Yano: Discovering Purchase Association among Brands from Purchase History, *Proceeding CD of SSGRR2002w* (2002).

[37] H.Hatazawa, H.Abe, M.Komori, Y.Tachibana and T.Yamaguchi: Knowledge Discovery Support from a Meningoencephalitis Dataset Using an AutomaticComposition Tool for Inductive Applications, *JSAI KDD Challenge 2001*, pp. 9-17 (2001).

[38] N.Izumi and T.Yamaguchi: Supporting Development of Software Agents by Integrating Heterogeneous Repositories Based on Ontologies, *5th International Conference on Autonomous Agents, WS on Ontologies and Agent Systems* (OAS2001)(http://CEUR-WS.org/Vol-52/), pp. 53-60,(2001).

[39] N.Izumi and T.Yamaguchi: Supporting Development of Business Applications Based on Ontologies, *International Conference on Enterprise Information Systems*, pp. 893-897 (2001).

[40] T.Yamaguchi: Acquiring Conceptual Relationships from Domain-Specific Texts, *17th International Joint Conference on Artificial Intelligence*, WS on Ontology Learning, pp. 14-19 (2001).

[41] N.Izumi and T.Yamaguchi: Building Business Applications by Integrating Heterogeneous Repositories Based on Ontologies, *17th International Joint Conference on Artificial Intelligence*, WS on Ontologies and Information Sharing, pp. 166-173 (2001).

[42] M.Kuremtsu, N,Nakaya and T.Yamaguchi: Acquiring Conceptual Relationships from a MRD and Text Corpus, *ECML/PKDD-2001*, WS on Semantic Web Mining,pp. 11-26 (2001).

[43] N.Izumi and T.Yamaguchi: Development Support of E-Commerce Applications by Integrating Heterogeneous Repositories Based on Ontologies, *International Conference on Electric Commerce 2001* (2001).

[44] Miho Ohsaki and Takahiro Sugiyama: A Research on Edge Detector Applications and Definition of Edge Quality, *1st International Symposium on Measurement*, Analysis and Modeling of Human Functions (ISHF'01), pp. 322-327 (2001).

[45] Miho Ohsaki, Shinich Sakamoto, and Hideyuki Takagi: Development and Evaluation of an IEC Fitting System for Hearing Aids, *17th International Congress on Acoustics* (ICA2001), 5A.15.02 (2001).

[46] E. Suzuki, M. Gotoh, and Y. Choki: Bloomy Decision Tree for Multi-Objective Classification, *Principles of Data Mining and Knowledge Discovery, Lecture Notes in Artificial Intelligence 2168 (PKDD)*, Springer-Verlag, pp. 436-447 (2001).

[47] E. Suzuki: Worst-Case Analysis of Rule Discovery, *Discovery Science, Lecture Notes in Artificial Intelligence 2226 (DS)*, Springer-Verlag, pp. 365-377 (2001). (erratum: http://www.slab.dnj.ynu.ac.jp/erratumds2001.pdf)

[48] N. Zhong, M. Ohshima, and S. Ohsuga: Peculiarity Oriented Mining and Its Application for Knowledge Discovery in Amino-acid Data, D. Cheung, G.J. Williams, Q. Li (Eds.), *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence 2035*, Springer-Verlag, pp. 260-269 (2001).

[49] J. Wu and N. Zhong: An Investigation on Human Multi-Perception Mechanism by Cooperatively Using Psychometrics and Data Mining Techniques, *Proc. 5th World Multi-Conference on Systemics, Cybernetics, and Informatics (SCI-01)*, in Invited Session on Multimedia Information: Managing and Processing, Vol. X, pp. 285-290 (2001).

[50] N. Zhong, Y.Y. Yao, M. Ohshima, and S. Ohsuga: Interestingness, Peculiarity, and Multi-Database Mining, *Proc. 2001 IEEE International Conference on Data Mining (IEEE ICDM'01)*, IEEE Computer Society Press, pp. 566-573 (2001).

[51] Taku Kudoh and Yuji Matsumoto: Chunking with Supprot Vector Machines, *Proceedings of the Second Meeting of North American Chapter of Association for Computational Linguistics* (NAACL), pp. 192-199, (2001).

[52] Tetsuji Nakagawa, Taku Kudoh, Yuji Matsumoto: Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines, *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium* (NLPRS2001), pp. 325-331 (2001).

[53] S. Tsumoto: Medical Knowledge Discovery in Hospital Information System, *Proceedings of SPIE Vol. 4384, Data Mining and Knowledge Discovery: Theory, Tools, and Technology III*, pp. 229-237 (2001).

[54] S. Tsumoto: Statistical Extension of Rough Set Rule Induction, *Proceedings of SPIE Vol. 4384, Data Mining and Knowledge Discovery: Theory, Tools, and Technology III*, pp. 180-188 (2001).

[55] S. Tsumoto: Mining Positive and Negative Knowledge in Clinical Databases Based on Rough Set Model, *Proceedings of the fifth European Conference on Principles of Knowledge Discovery in Databases* (PKDD2001), pp. 460-471 (2001).

[56] S. Hirano and S. Tsumoto: A Knowledge-Oriented Clustering Technique Based on Rough Sets, *Proceedings of the 25th IEEE International Computer Software and Applications Conference* (Compsac2001), pp. 632-637 (2001).

[57] S. Hirano and S. Tsumoto: Indiscernability Degrees of Objects for Evaluating Simplicity of Knowledge in the Clustering Procedure, *Proceedings of the 2001 IEEE International Conference on Data Mining*, pp. 211-217 (2001).

[58] S. Tsumoto: Temporal Knowledge Discovery in Time-Series Medical Databases based on Fuzzy and Rough Reasoning, *Proceedings of Ninth International Fuzzy Systems Association World Congress* (IFSA'01), (CD-ROM) (2001).

[59] S. Tsumoto: Discovery of Temporal Knowledge in Medical Time-Series Databases Using Moving Average, Multiscale Matching, and Rule Induction, *Proceedings of the fifth European Conference on Principles of Knowledge Discovery in Databases* (PKDD2001), pp. 448-459 (2001).

[60] S. Hirano and S. Tsumoto: Analysis of Time-series Medical Databases Using Multiscale Structure Matching and Rough Sets-based Clustering Technique, *Proceedings of the 2001 IEEE International Conference on Fuzzy Systems* (2001).

[61] S. Tsumoto, S. Hirano, A. Yasuda, and K. Tsumoto: Analysis of Amino Acid Sequences by Statistical Technique, *Proceedings of the 4th Conference on Computational Biology and Genome Informatics* (CBGI-02), 2002 (in press).

[62] T. Okada: Characteristic Substructures and Properties in the Chemical Carcinogenicity Studied by the Cascade Model, *International workshop on Predictive Toxicology Challenge 2001*, Freiburg (2001).

[63] T. Niwa, K. Fujikawa, K. Tanaka, and M. Oyama: Visual Data Mining Using a Constellation Graph, *International Workshop on Visual Data Mining*, Freiburg (2001).

[64] Y. Takahashi, and M. Konji: Visualization of Massive Data in Molecular Space and Similar Structure Searching, *6th China-Japan Joint Symposium on Drug Design and Development*, pp. 1-3, Dalian, China (2001).

[65] H. Kato, Y. Takahashi, and H. Abe: Development of Automated Identification System for Three-dimensional Protein Motifs, *10th German-Japanese Workshop on Chemical Information*, pp. 6, Potsdam, Germany (2001).

[66] S. Yamada and Y. Ohsawa: Information Gathering of Web pages to Guide Concept Understanding *Posters Proc. the World Wide Web Conference (WWW10)* Hong Kong, (May 2001)

[67] Y. Ohsawa, N. Matsumura and M. Ishizuka: Discovering Topics to Enhance Communities' Creation from Links to the Future, *Posters Proc. the World Wide Web Conference (WWW10)* Hong Kong, (May 2001)

[68] N. Matsumura and Y. Ohsawa: Future Deirections of Communities on the Web, *The First Workshop on Chance Discoveries, Japanese Society of Artificial Intelligence*Matsue, Japan (May 2001)

[69] H. Taira, Y. Sakata, Y. Ohsawa and M. Ishizuka: AreaView2001: A new WWW organization system with KeyGraph Technology *The First Workshop on Chance Discoveries, Japanese Society of Artificial Intelligence*Matsue, Japan (May 2001)

[70] Y. Matsuo and Y. Ohsawa: A Document as a Small World, *The First Workshop on Chance Discoveries, Japanese Society of Artificial Intelligence*Matsue, Japan (May 2001)

[71] Y. Nara and Y. Ohsawa: A Method for Discovering Seeds of Consensus Applied to Family Risk Perceptions *The First Workshop on Chance Discoveries, Japanese Society of Artificial Intelligence*Matsue, Japan (May 2001)

[72] Y. Matsuo, Y. Ohsawa and M. Ishizuka: Small-World as Asserting Structure of Document, *Proc. The Fifth Multi-Conference on Systems, Cybernetics and Informatics (SCI2001)*, Orland, Florida USA (August 2001)

[73] N. Matsumura, Y. Ohsawa and M. Ishizuka: Future Directions of Communities on the Web, *Proc. The Fifth Multi-Conference on Systems, Cybernetics and Informatics (SCI2001)*, Orland, Florida USA (August 2001)

[74] Y. Ohsawa and Y. Nara: Family Perceptions of Risks and Opportunities - Results from Questionnaires to Citizens -, The Fifth Multi-Conference on Systems, Cybernetics and Informatics (SCI2001), Orland, Florida USA (August 2001)

[75] Y. Nara and Y. Ohsawa: Family Affection to Children Visualized on the Co-occurring of Questionnaire Answers, *Proc. The Fifth Multi-Conference on Systems, Cybernetics and Informatics (SCI2001)*, Orland, Florida USA (August 2001)

[76] F. Yoshikawa and Y. Ohsawa: Text Analysis by Mapping Information Flow, *Proc. The Fifth Multi-Conference on Systems, Cybernetics and Informatics (SCI2001)*, Orland, Florida USA (August 2001)

[77] Y. OHSAWA and Y. NARA: Decision Trees as a Model of Chance Perception, *Joint 9th IFSA Congress and 20th NAFIPS International Conference*, Vancouver (August 2001)

[78] H. Fukuda and Y. OHSAWA: Discovery of Rare Essential Food by Community Navigation with KeyGraph - An introduction to Data-based Community Marketing - *Proc. KES 2001 (from IOS press)* (September 2001)

[79] N. MATSUMURA, Y. OHSAWA and M. ISHIZUKA: Discovery of Emerging Topics by Co-citation Graph on the Web *Proc. KES 2001 (from IOS press)* (September 2001)

[80] Y. Matsuo, Y. OHSAWA and M. ISHIZUKA: Discovering Hidden Relation behind a Link *Proc. KES 2001 (from IOS press)* (September 2001)

[81] Y. OHSAWA and Y. NARA: Discovery of Virtual Behaviors as Signs of Real Behaviors *Proc. KES 2001 (from IOS press)* (September 2001)

[82] Y. OHSAWA and N. MATSUMURA: Discovering Seeds of New Interest Spread from Premature Pages Cited by Multiple Communities *Proc. Web Intelligence (LNAI2198 from Springer Verlag)* (October 2001)

[83] N. MATSUMURA, Y. OHSAWA and M. Ishizuka: Discovery of Emerging Topics between Communities on WWW *Proc. Web Intelligence (LNAI2198 from Springer Verlag)* (October 2001)

[84] Y. Matsuo, Y. OHSAWA and M. Ishizuka: Average-clicks: A New Measure of Distance on the World Wide Web *Proc. Web Intelligence (LNAI2198 from Springer Verlag)* (October 2001)

[85] Y. Matsuo, Y. OHSAWA and M. Ishizuka: KeyWorld: Extracting Keywords from a Document as a Small World, *Proc. The Forth International Conference on Discovery Science (An LNAI from Springer Verlag)* (Washington DC, December 2001)

[86] N. Matsumura, Y. OHSAWA and M. Ishizuka: Knowledge Navigation on Visualizing Complementary Documents *Proc. The Forth International Conference on Discovery Science (An LNAI from Springer Verlag)* (Washington DC, December 2001)

[87] S. Kurahashi, T. Terano: Can We Control Information Free Riders? Analyzing Communal Sharing Norms via Agent-based Simulation. (CASOS 2001), pp. 94-96 (2001).

[88] H. Matsui, I. Ono, H. Sato, H. Deguchi, T. Terano, H. Kita, Y. Shinozawa: Learning Economics Principles from Bottom, (CASOS 2001), pp. 97-99 (2001).

[89] S. Kurahashi, T. Terano: Analyzing Norm Emergence in Communal Sharing via Agent-based Simulation, (AESCS-2001), pp. 27-34 (2001).

[90] H. Deguchi, T. Terano, K. Kuramarani, T. Yuzawa, S. Hashimoto, H. Matsui, A.Sashima, T.Kaneda: Virtual Economy Simulation and Gaming -An Agent Based Approach-, (AESCS-2001), pp. 169-185 (2001).

[91] Y. Katsumata, A, S. Kurahashi, T. Terano: Hybridizing Bayesian Optimization and Tabu Search for Multimodel Fancrions, *2001 Genetic and Evolutionary Computation Conference* (GECCO-2001 Late Breaking Papers, pp. 227-333, 2001

[92] K. Taniguchi, S. Kurahashi, T. Terano: Managing Information Complexity in a Supply Chain Model Agent-Based Genetic Programming, *2001 Genetic and Evolutionary Computation Conference* (GECCO-2001) Late Breaking Papers, pp. 413-420, 2001

[93] E. Murakami, T. Terano: Collaborative Filtering for a Distributed Smart IC Card System. Intelligent Agents: Specification, Modeling, and Applicatons, *4th Pacific Rim International Workshop on Muli-Agents*, PRIMA 2001, Taipai, Taiwan, July 2001, Proceedings, Springer LNAI2132, pp. 183-197, 2001

[94] T. Terano, Y. Shiozawa, H. Deguchi, H. KIta, H. Matsui, H. Sato, I. Ono: U-Mart: An Artificial Market to Bridge the Studies on Economics and Multiagent Systems, *Proc. PRIMA'2001, 4th Pacific Rim International Workshop on Multi-agents*, pp. 371-385, Taipei, July 28-29, 2001

[95] K. Takadama, T. Terano, K. Shimohara: Learning Classifier Systems Meet Multi-agent Environments, Pier Luca Lanzi,Wolfgang Stolzmann,Stewaer W.Wilson(Eds.): *Advances in Learning Classifier Sysems*, Third International Workshop, IWLCS 2000, Paris, France, September 2000. Springer LNAI 1996, pp192-212, 2001

3.3 国内発表

[1] 山田誠二，岡部正幸：関係学習を用いた対話的 Web ページフィルタリング, 第 54 回人工知能学会「知識ベースシステム」研究会, pp. 67-72 (2001).

[2] 中井有紀，山田誠二：Web ページにおける部分情報の更新モニタリング, 第 54 回「知識ベースシステム」研究会，pp. 73-78 (2001).

[3] 岡部正幸，山田誠二：関係学習を用いたフィルタ生成による対話的 Web ページ検索, 第 61 回情報学基礎研究会，pp. 197-204 (2001).

[4] 高間康史, 廣田薫：

可塑的クラスタリングに基づく WWW 情報可視化システム, 電気学会システム・制御研究会, SC-01-15, pp. 1-6 (2001).

[5] 高間康史, 廣田薫：免疫ネットワークを用いたキーワード集合抽出の情報可視化システムへの応用, 第 17 回ファジィシステムシンポジウム（日本ファジィ学会）, pp. 781-782 (2001).

[6] 高間康史, 廣田薫: WWW 上の情報収集／可視化のための免疫ネットワークを用いたクラスタリング, 第 46 回人工知能基礎論研究会／第 54 回知識ベースシステム研究会（合同研究会）（人工知能学会）, pp. 61-66 (2001).

[7] 北村泰彦：Web とエージェントと教育システム．人工知能学会第 16 回 AI シンポジウム「e-learning の intelligent 化に向けて」, 人工知能学会研究会資料, SIG-J-A102-9, pp. 47-48 (2001).

[8] 北村泰彦：WWW 情報統合からエージェント統合へ, 情報処理学会連続セミナー 2001「21 世紀のネットサービス社会」第 4 回「サービスぷプラットフォーム技術」, pp. 1-2 (2001).

[9] 北村泰彦：アクティブ情報収集システムに関する検討, 人工知能学会人工知能基礎論・知識ベースシステム研究会, SIG-FAI/KBS-J-14 (2001).

[10] 辻本秀樹，北村泰彦，辰巳昭治：

Wizard of Oz 法を使ったマルチキャラクタエージェントインタフェースの評価, 日本ソフトウェア科学会第 10 回マルチエージェントと協調計算ワークショップ (2001).

[11] 阪本俊樹，北村泰彦，辰巳昭治：競争型推薦システム Recommendation Battlers とその挙動, 人工知能学会人工知能基礎論・知識ベースシステム研究会，SIG-FAI/KBS-J-09 (2001).

[12] 植村渉，辰巳昭治，北村泰彦: 強化学習を用いた 2D メッシュ結合型マルチコンピュータでの耐故障性を持つ適応経路設定, 電子情報通信学会情報・システムソサイエティ大会、D-10-7 (2001).

[13] 山田晃弘，小久保卓，北村泰彦: マルチキャラクタインタフェースを用いた Web 情報統合, 人工知能学会第 15 回全国大会，1F1-05 (2001).

[14] 阪本俊樹，回り道康博，北村泰彦，辰巳昭治: マルチキャラクタを用いた競争型情報推薦システム, 人工知能学会第 15 回全国大会，1F1-01 (2001).

[15] 阪本俊樹，回り道康博，北村泰彦，辰巳昭治: マルチエージェントによる競争型情報推薦システム, 電子情報通信学会関西支部第 6 回学生会研究発表講演会, D-1 (2001).

[16] 河野浩之: 位置データ問合せ処理のための空間インデックス手法の検討, 第 45 回システム制御情報学会研究発表講演会 (2001).

[17] 中辻 真，川原稔，河野浩之: ピアツーピアネットワークにおけるトピック主導型検索手法の提案, 知能と複雑系研究会, pp. 47-54 (2001).

[18] 南 卓朗, 田名部 淳, 河野 浩之: 空間インデックスを用いた移動オブジェクト管理システムの構成と性能比較, 情処研報 Vol. 2001, No. 70, DBS-125, pp. 225-232 (2001).

[19] 河野浩之: 位置情報システムにおける空間データ利用に関する検討, 人工知能学会 FAI-46&KBS-54 研究会, pp. 159-164 (2001).

[20] 乾岳史，櫻井成一朗: マルチエージェントによる組織形成に関する基礎的研究, 情報処理学会, 予稿 (2001-GI-5)，pp. 23-30 (2001).

[21] 白壁啓吾，乾岳史，櫻井成一朗: マルチエージェント系における強化結果に対する解釈について, 情報処理学会, 予稿 (2001-ICS-124)，pp. 55-60 (2001).

[22] 若月謙太郎，櫻井　成一朗: リンク解析による WWW ページ群の発見, 人工知能学会, 予稿 (2001-FAI-46, 2001-KBS-54)，pp. 109-114 (2001).

[23] 沼尾正行: Global intelligence による知識流通, 情報処理学会知能と複雑系研究会, Vol. 01-ICS-124, pp. 1-8 (2001).

[24] ナッティー・チョラウィト, 沼尾正行: 知識流通における紙面レイアウトの役割とそれに基づく自動タグ 付け, 情報処理学会知能と複雑系研究会, Vol. 01-ICS-124, pp. 17-24 (2001).

[25] 伊藤雄介, 吉田匡史, 沼尾正行: 口コミ支援システムの実験, 情報処理学会知能と複雑系研究会, Vol. 01-ICS-124, pp. 9-16 (2001).

[26] 老川正志, 沼尾正行: 帰納論理プログラミングによる嗜好情報の獲得とナビゲーションシステム, 人工知能学会全国大会（第 15 回）論文集 (2001).

[27] 高木将一, 中村啓佑, 沼尾正行: 機械学習の手法を用いた感性の抽出と作曲・編曲への応用, 人工知能学会全国大会（第 15 回）論文集 (2001).

[28] 五十嵐建平, 大田佳宏, 横山茂樹, 沼尾正行: データマイニングにおける XML を用いたデータ構造の変形, 人工知能学会全国大会（第 15 回）論文集 (2001).

[29] 伊藤雄介, 吉田匡史, 沼尾正行: 多くの人の評価を経て情報が吟味される口コミ支援システム, 人工知能学会全国大会（第 15 回）論文集 (2001).

[30] Cholwich Nattee and Masayuki Numao: Document analysis and recognition using ILP and the winnow algorithm, 人工知能学会全国大会（第 15 回）論文集 (2001).

[31] 沼尾正行, 高木将一, 中村啓佑: ユーザの感性に合わせた自動編曲及び作曲, 情報処理学会音楽情報科学研究会, Vol. 2001-MUS-41, pp. 49-54 (2001).

[32] 森山甲一, 沼尾正行: 自己の報酬を操作する学習エージェントの構築, 人工知能学会第 45 回人工知能基礎論研究会, 第 SIG-FAI-A101 巻, pp. 15-20, (2001).

[33] 森山甲一, 沼尾正行: 環境状況の変化に応じて自己の報酬を操作する学習エージェントの構築, http://www-kasm.nii.ac.jp/macc2001-proceedings/MACC2001-15.pdf. MACC2001, (2001).

[34] 吉田匡志, 伊藤雄介, 沼尾正行: 口コミによる分散型情報収集システム – WAVE を起こそう – Word-of-mouth-Assisting Virtual Environment, http://www-kasm.nii.ac.jp/macc2001-proceedings/MACC2001-10.pdf, MACC2001 (2001).

[35] 沼尾正行, 吉田匡志, 伊藤雄介: 口コミに基づく情報収集とデータ前処理, 人工知能学会第 46 回人工知能基礎論/第 54 回知識ベースシステム研究会, 第 SIG-FAI/KBS-J 巻, pp. 47-54 (2001).

[36] Tuan Nam Tran and Masayuki Numao: Mining biomedical literature by making use of existing databases, 人工知能学会第 46 回人工知能基礎論/第 54 回知識ベースシステム研究会, 第 SIG-FAI/KBS-J 巻, pp. 41-46 (2001).

[37] 西村 芳男, 鷲尾　隆, 元田　浩, 猪口 明博: 大量データからの誘導部分グラフデータの検索手法, 2001 年度人工知能学会全国大会資料 (第１５回）, 2D1-04 pp. 1-4 (2001).

[38] 寺邊　正大, 鷲尾　隆, 元田　浩: 相関ルールに基づく属性生成手法 －連続値属性を含むデータへの適用－, 2001 年度人工知能学会全国大会資料 (第１５回）, 2D1-09, pp. 1-4 (2001).

[39] 和田卓也, 元田　浩, 鷲尾　隆: クラス分布の変化に追従するための Ripple Down Rules 法の拡張に関する実験, 2001 年度人工知能学会全国大会資料 (第１５回）, 1B2-03, pp. 1-4 (2001).

[40] 鷲尾　隆, 元田　浩, 丹羽 雄二: 観測データからの第一原理に基づく連立方程式の発見, 2001 年度人工知能学会全国大会資料 (第１５回）, 2D1-03, pp. 1-4 (2001).

[41] 藤原 啓成, 元田　浩, 鷲尾　隆: 専門家の事例判断のみを利用した RDR 知識ベース構築のための事例生成手法の評価実験, 第 46 回人工知能基礎論研究会, 第 54 回知識ベースシステム研究会（合同研究会）資料 (SIG-FAI/KBS-J-25), pp. 153-158 (2001).

[42] 和田卓也, 元田　浩, 鷲尾　隆: 環境変化への適応, 異種の知識源からの知識獲得を目的とした Ripple Down Rules 法の拡張に対する評価実験, 第 46 回人工知能基礎論研究会, 第 54 回知識ベースシステム研究会（合同研究会）資料 (SIG-FAI/KBS-J-26), pp. 165-170 (2001).

[43] 松田　喬, 元田　浩, 鷲尾　隆: Graph-Based Induction の部分グラフ抽出能力の改良, 第 46 回人工知能基礎論研究会, 第 54 回知識ベースシステム研究会（合同研究会）資料 (SIG-FAI/KBS-J-30), pp. 185-187 (2001).

[44] T.B. Ho, S. Kawasaki, and D.D. Nguyen: Extracting Predictive Knowledge from Meningitis Data by Integration of Rule Induction and Association Mining, *Proc. of International Workshop Challenge in KDD, JSAI Annual Conference*, pp. 25-32 (2001).

[45] T.B. Ho, T.D. Nguyen, D.D. Nguyen, and S. Kawasaki: Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining, 人工知能基礎論研究会, SIG-FAI/KBS-J-35, pp. 207-212 (2001).

[46] 矢田勝俊: データマイニング技術とカスタマープロファイリング, 商業学会関西部会 (2001).

[47] 矢田勝俊, 飯田洋: ビジネス知識発見の研究展望とアクティブマイニング, 人工知能基礎論研究会, SIG-FAI/KBS-J-35, pp. 213-217 (2001).

[48] 加藤直樹, 羽室行信: 矢田勝俊購買履歴からのデータマイニング, 2001 年情報論的学習理論ワークショップ論文集, pp. 95-104 (2001).

[49] 小森麻央, 阿部秀尚, 畑澤寛光, 橘恵昭, 山口高平: 知識発見システムにおける属性処理と属性値処理に関する一考察, 第１５回人工知能学会全国大会, 1F1-08 (2001).

[50] 中矢尚美, 岩出 剛昌, 樗松理樹, 山口高平: 情報リソースを活用した領域オントロジー構築支援環境, 第１５回人工知能学会全国大会, 2B2-03 (2001).

[51] 杉浦直樹, 河口 知幸, 和泉憲明, 山口高平：ビジネスリポジトリに基づくビジネスモデル構築支援環境の開発, 第１５回人工知能学会全国大会, 2F1-05 (2001).

[52] 阿部秀尚, 山口高平: 共通データによる帰納アプリケーション自動構築支援環境の評価, 人工知能学会, 第４６回人工知能基礎論研究会・第５４回知識ベースシステム研究会合同研究会 , pp. 195-200 (2001).

[53] 和泉憲明, 杉浦直樹, 澤井雅彦, 寺井公一, 山口高平：ビジネスアプリケーション開発のための異種リポジトリの統合法, 電子情報通信学会, ソフトウェアインタプライズモデリング研究会, pp. 9-16 (2002).

[54] 和泉憲明, 杉浦直樹, 澤井雅彦, 寺井公一, 山口高平：ビジネスアプリケーション開発のための多粒度リポジトリの開発, 人工知能学会, 第５５回知識ベースシステム研究会, pp. 91-98 (2002).

[55] 山口高平: 金融データにおけるＤＭアプリケーションの自動合成, 人工知能学会第 17 回ＡＩシンポジウム (2002).

[56] 大崎美穂, 坂本真一, 高木英行: IEC フィッティングの実用性の評価, 日本ファジィ学会東海支部主催, 第 11 回東海ファジィ研究会, pp. 1.1-1.4 (2001).

[57] 藤井成清, 林田憲昌, 高木英行, 大崎美穂: PDA 版 Visualized IEC フィッティングシステム, 第 3 回日本ファジィ学会九州支部学術後援会, pp. 61-64 (2001).

[58] 鈴木英之進: ルール発見の最悪解析, 第 46 回人工知能学会人工知能基礎論研究会& 第 54 回人工知能学会知識ベースシステム研究会 合同研究会, pp. 189-194 (2001).

[59] 長木悠太, 鈴木英之進: 反復マハラノビスデータ圧縮に基づく高速ブースティング, 第 46 回人工知能学会人工知能基礎論研究会 & 第 54 回人工知能学会知識ベースシステム研究会 合同研究会, pp. 201-206 (2001).

[60] 長浜光俊, 山口直記, 鈴木英之進：粗利と購買履歴に基づく有望顧客の特定, ビジネスマイニングワークショップ講演論文集, pp. 20-23 (2001).

[61] 長木悠太, 鈴木英之進：反復データ圧縮型ブースティングの実験的評価, 第 48 回人工知能学会人工知能基礎論研究会 (2002).

[62] 中本和岐, 鈴木英之進: ＴＷＳ木を用いた例数圧縮による時系列データの高速クラスタリング, 第 48 回人工知能学会人工知能基礎論研究会 (2002).

[63] 武智文雄, 鈴木英之進: 集合属性の利得比上限値に基づく決定木の高速学習, 第 48 回人工知能学会人工知能基礎論研究会 (2002).

[64] 鈴木英之進：データマイニングにおけるデータ変換, 人工知能学会第 17 回 AI シンポジウム (2002).

[65] 鈴木英之進：データマイニングにおける例外逸脱発見, 統計数理とデータマイニング・発見科学 研究会 (2002).

[66] N. Zhong and M. Ohshima: Mining Interesting Patterns in Multiple Data Sources, *Research Report of JSAI SIG-FAI/KBS-J-01*, Hakodate, Japan, No. 12-14, pp. 177-184 (2001).

[67] 松本裕治，工藤拓，高村大也，山田寛康，中川哲治: 自然言語処理におけるシステム混合法について第 4 回情報論的学習理論ワークショップ予稿集, pp. 19-24 (2001).

[68] 松本裕治，山田寛康，新保 仁: 学習に基づく専門用語分類, 人工知能学会・人工知能基礎論研究会・知識ベースシステム研究会合同研究会 SIG-FAI/KB S-J-13, pp. 79-84 (2001).

[69] 中川哲治，工藤拓，松本裕治: 修正学習法による形態素解析, 情報処理学会研究報告 2001-NL-146, pp. 1-8 (2001).

[70] 山田寛康，松本裕治: Support Vector Machine の多値分類問題への適用法について, 情報処理学会研究報告 2001-NL-146, pp. 33-38 (2001).

[71] 津本周作: 医療情報から見た歯科医療界の今後-最新の情報技術と診療支援-, 島根県歯科医師会生涯教育講座, pp. 1-12 (2001).

[72] 津本周作: 特別講演: 医療におけるアクティブマイニング, -Medical Data Mining からの新たな展開-, バイオメディカルファジイシステム学会第 14 回年次大会講演論文集, pp. 1-4 (2001).

[73] 平野章二, 津本周作: ラフ集合論に基づく知識指向型クラスタリング法, バイオメディカルファジイシステム学会第 14 回年次大会講演論文集, pp. 6-9 (2001).

[74] 平野章二, 津本周作: 多重スケールマッチングとラフクラスタリングによる時系列臨床検査データベースの解析, 人工知能学会第 46 回人工知能基礎論研究会, 第 54 回知識ベースシステム研究会合同研究会資料 (SIG-FAI/KBS-J-42), pp. 257-260 (2001).

[75] 孫 暁光, 柳樂真佐実, 平野章二, 安田 晃, 津本周作: 臨床データベース解析のための類似性尺度とその評価, 第 21 回医療情報学連合大会講演論文集, pp. 504-505 (2001).

[76] 岡田孝：Challenge における化学物質による発ガン性の予測, 第 24 回情報化学討論会, 第 29 回構造活性相関シンポジウム, JK06, pp. 13-14 (2001).

[77] 岡田孝: アクティブマイニングのためのルール群表現法, 人工知能学会, 第 46 回基礎論研究会 (SIGFAI) 第 54 回知識ベースシステム研究会 (SIG-KBS) 合同研究会，pp. 219-224, 函館 (2001).

[78] 岡田孝: カスケードモデルによる高次元データの解析: 化学物質の構造活性相関を対象として, 研究集会「高次元データ解析の研究」, 広島 (2002).

[79] 高橋由雅, 加藤博明, 藤島悟志: 化学物質の構造類似性にもとづくデータマイニング, 人工知能学会第 46 回基礎論研究会 (SIGFAI) 第 54 回知識ベースシステム研究会 (SIG-KBS) 合同研究会, pp. 225-228, 函館 (2001).

[80] 藤島悟志, 高橋由雅: トポロジカルフラグメントスペクトル (TFS) ピーク同定システムの開発, 第 24 回情報化学討論会, J15, pp. 59-60 (2001).

[81] 加藤博明, 田所哲男, 宮田博之, 高橋由雅, 阿部英次: タンパク質三次元構造モチーフデータベースの作成, 第 24 回情報化学討論会, JP14, pp. 125-126 (2001).

[82] 石原由一朗, 高橋由雅: 構造類似性を基礎とした化学物質の毒性予測のシステム化に関する研究, 第 29 回構造活性相関シンポジウム, K14, pp. 213-214 (2001).

[83] 松村真宏, 大澤幸生, 石塚満：語の活性度に基づくキーワード抽出法, 第 2 回人工知能学会 MYCOM (2001).

[84] 松尾 豊, 大澤 幸生, 石塚 満：Small world と Average-clicks, 第 2 回人工知能学会 MYCOM (2001)

[85] 大澤幸生, 松村真宏, 松尾豊: 自然・社会現象データからの予兆・チャンス発見, 第 43 回人工知能学会 AI チャレンジ研究会 (SIG-Challenge)(2001)

[86] 大澤幸生, 相馬浩隆, 臼井優樹, 松村真宏, 松尾豊：会話展開キーグラフによる Web コミュニティの特性表現, 人工知能学会第 46 回 SIG-FAI 第 54 回 SIG-KBS 合同研究会 (2001)

[87] 福田寿・大澤幸生: 消費者の食品メニュー履歴データからの潜在意識発見支援, 人工知能学会・第 47 回人工知能基礎論研究会 (2002)

[88] 松村真宏・大澤幸生・石塚満: 頻出アイテムと希少アイテム間のコレレーションルールからのチャンス発見, 人工知能学会・第 47 回人工知能基礎論研究会 (2002)

[89] 松村真宏・大澤幸生・石塚満: 語の活性伝播に基づく談話分析, 人工知能学会・第 47 回人工知能基礎論研究会 (2002)

[90] 大澤幸生・奈良由美子: チャンス発見プロセスの二重らせんモデルに基づくアンケート調査データの解析, 人工知能学会・第 47 回人工知能基礎論研究会 (2002)

[91] 吉川史子・大澤幸生: 新聞記事に見るチャンスの意味, 人工知能学会・第 47 回人工知能基礎論研究会 (2002)

[92] 松尾豊・大澤幸生・石塚満: 電子掲示板における会話からのハイライト部分の抽出, 人工知能学会・第 47 回人工知能基礎論研究会 (2002)

[93] 松尾豊・大澤幸生・石塚満: Small World のさまざまな拡張についての考察, 人工知能学会第 47 回人工知能基礎論研究会 (2002)

[94] 松村真宏・大澤幸生・石塚満：Knowledge Navigation through Combining Documents, 人工知能学会全国大会, 1A4-06 (2001)

[95] 坂田恭弘・平博司・大澤幸生・伊庭斉志, 石塚満: AreaBook: WWW エリアビューのブック型情報提示インタフェース, 人工知能学会全国大会, 1A4-06 (2001)

[96] 松村真宏、大澤幸生、石塚満：質問応答システムのための新しい文書検索手法の提案: 第 40 回情報処理学会全国大会 (2001)

[97] 松村真宏、大澤幸生、石塚満：文書の組み合わせに基づく知識ナビゲーション: 第 15 回人工知能学会全国大会 (2001)

[98] 第 47 回人工知能基礎論研究会 (SIG-FAI47) パネルディスカッション「アクティブマイニング時代におけるチャンス発見の役割」(2002)

[99] (財) 日本科学技術連盟・多変量解析シンポジウム招待講演「人のチャンス発見プロセスにおける情報視覚化と KeyGraph」(2002)

[100]大澤幸生: 近未来チャレンジのきっかけと展開, 第 43 回人工知能学会 AI チャレンジ研究会 (SIG-Challenge)(2001)

[101]勝又勇治、倉橋節也、寺野隆雄：ベイジアンネットワークとタブーリストを利用したハイブリッドＧＡによる多峰性関数の最適化, 人工知能学会全国大会 (第 15 回) 論文集, pp. 2C3-01 (2001).

[102]谷口 憲 , 倉橋 節也 , 寺野 隆雄：エージェント指向サプライ・チェーン・モデルに対する遺伝的プログラミングの適用, 人工知能学会全国大会 (第 15 回) 論文集, pp. 2C3-09 (2001).

[103]倉橋 節也 , 寺野 隆雄：エージェントモデルによるコミュニティの共同分配規範の成立と崩壊に関する考察, 人工知能学会全国大会 (第 15 回) 論文集, pp. 3C2-03 (2001).

[104]村上 英次 , 寺野 隆雄：F-CUBE：コミュニティに共有する知識のインターネットフォーラムからの抽出システム, 人工知能学会全国大会 (第 15 回) 論文集, pp. 3B1-02 (2001).

[105]出口 弘 , 寺野 隆雄 , 車谷 浩一 , 湯澤 太郎 , 橋本 重治 , 松井 啓之 , 幸島 明男 , 兼田 敏之：エージェントに基づく仮想経済シミュレータの開発, 人工知能学会全国大会 (第 15 回) 論文集, pp. 2F1-10 (2001).

[106]佐藤 浩 , 松井 啓之 , 小野 功 , 喜多 一 , 寺野 隆雄 , 出口 弘 , 塩沢 由典：U-Mart: エージェントシミュレーションで経済を学ぶ, 人工知能学会全国大会 (第 15 回) 論文集, pp. 3F1-11 (2001).

[107]山口 高平 , 元田 浩 , 寺野 隆雄 , 鷲尾 隆 , 斉藤 和巳 , 津本 周作：Discovery of Communicable Knowledge, 人工知能学会全国大会 (第 15 回) 論文集, pp. 3B4-01 (2001).

[108]勝又勇治，倉橋節也，寺野隆雄：タブーリストを用いたベイジアン最適化アルゴリズムによる多峰性関数最適化，情報処理学会　第 8 回 MPS（数理と問題解決) シンポジウム-進化的計算シンポジウム 2001-，2001

[109]寺野隆雄：計算組織論とエージェントベースモデル, 社会・経済システム学会　第 20 回大会プログラム「システム論を問いなおす：システム論の新展開-主題と方法-」，(2001).11.10-11

[110]寺野隆雄，稲田政則：EBM とデータマイニング：知識評価の観点から, 人工知能学会研究会資料　SIG-FAI/KBS-J-16(11/13)，pp97-102，2001

[111]高橋大志，寺野隆雄:エージェントシミュレーションによる Prospect 理論と GARCH モデルの関連性の分析, 第 10 回マルチ･エージェントと協調計算ワークショップ (MACC2001) (2001)

[112]喜多一，出口弘，寺野隆雄：U-Mart：経済学と工学をエージェントが結ぶ，第 10 回マルチ・エージェントと協調計算ワークショップ (MACC2001) (2001)

[113]國上真章，寺野隆雄：エージェント系としてのレプリケータダイナミックスとその制御，第 10 回マルチ・エージェントと協調計算ワークショップ (MACC2001) (2001).

[114]稲田政則，寺野隆雄：Evidence-Based Medicine とデータマイニング，計測自動制御学会，システム・情報部門学術講演会 (2001).

[115]喜多一，出口弘，寺野隆雄：オープン型人工市場 U-Mart：構想，成果，展望，電子情報通信学会技術研究報告，人工知能と知識処理（AI2001 － 58)，pp. 17-23，(2002).

[116]寺野隆雄，出口弘：社会科学におけるエージェント研究の動向と課題，電子情報通信学会技術研究報告，人工知能と知識処理（AI2001 － 59)，pp25-32，(2002).

[117]寺野隆雄，喜多一，出口弘：U-Mart プロジェクト：仮想市場とエージェントと経済学，人工知能学会，知識ベースシステム研究会（第 55 回)，pp99-103，(2002)

3.4　著書

[1] 北村泰彦:探索，教育システム情報ハンドブック，実教出版, ISBN4-407-05118-3 (2001).

[2] N. Zhong: Knowledge Discovery and Data Mining, *Encyclopedia of Microcomputers*, Vol. 27 (Supplement 6), Marcel Dekker, pp. 235-286 (2001).

[3] N. Zhong and S. Ohsuga: Automatic Knowledge Discovery in Larger Scale Knowledge-Data Bases, in C. Leondes (ed.) *The Handbook of Expert Systems*, Vol. 4, Academic Press, pp. 1015-1070 (2001).

[4] H. Liu and H. Motoda: *Instance Selection and Construction for Data Mining* (Eds.), Kluwer Academic Publishers (2001).

[5] L. Polkowski, S. Tsumoto, and T.Y. Lin (eds.): *Rough Set Methods and Applications : New Developments in Knowledge Discovery in Information Systems*, Physica-Verlag, New York (2001).

[6] S. Tsumoto: Discovery of Clinical Knowledge in Databases Extracted from Hospital Information Systems, K.J. Cios (ed.) *Medical Data Mining and Knowledge Discovery*, Physica-Verlag, New York, pp. 433-454 (2001).

[7] S. Tsumoto: Induction of Rule about Complications with the Use of Rough Sets, W. Pedrycz (ed.) *Granular Computing: an emerging paradigm*, Physica-Verlag, New York, pp. 384-397 (2001).

[8] S. Tsumoto: Chapter G5: Data Mining in Medicine, W. Kloesgen and J. Zytkow (eds.) *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press (2001).

[9] T. Terano, T. Nishida, A. Namatame, S. Tsumoto, Y. Ohsawa, and T. Washio (eds.): *New Frontiers in Artificial Intelligence*, Joint JSAI 2001 Workshop Post-Proceedings, Lecture Notes in Computer Science 2253, Springer-Verlag, Heidelberg (2001).

[10] N. Callaos, Y. Ohsawa, Y. Zhang, R. Szabo, and M. Aveledo, eds.: *World Multiconference on Systemics, Cybernetics, and Informatics Proceedings Volume VIII* (2001)

[11] Y. Ohsawa: Discovery of Chances Underlying Real Data, *Progress in Discovery Science*, eds., Arikawa, S., LNAI from Springer Verlag (2001)

[12] T. Terano, T. Nishida, A. Namatame, S. Tsumoto, Y. Ohsawa, and T. Washio, eds.: *New Frontiers in Artificial Intelligence, Joint JSAI2001 Workshops Post-Proceedings* LNAI2253, Springer Verlag (2001)

[13] Y. Ohsawa, The Scope of Chance Discovery: *New Frontiers in Artificial Intelligence,* LNAI2253, T. Terano, et al eds., Springer Verlag, pp. 413(2001)

[14] N, Matsumura, Y. Ohsawa and M. Ishizuka: Chapter 59: Future Directions of Communities on the Web *New Frontiers in Artificial Intelligence,* LNAI2253, T. Terano, et al eds., Springer Verlag, pp. 435-442(2001)

[15] Y. Ohsawa and Y. Nara: Chapter 66: Action Proposals as Discovery of Context (An Application to Family Risk Management) *New Frontiers in Artificial Intelligence,* LNAI2253, T. Terano, et al eds., Springer Verlag, pp. 481-485 (2001)

[16] Y. Matsuo and Y. Ohsawa: Chapter 60: A Document as a Small World *New Frontiers in Artificial Intelligence,* LNAI2253, T. Terano, et al eds., Springer Verlag, pp. 444-448 (2001)

[17] K.TAKADAMA, T. Terano, K.SHIMOHARA, K.HORI, S.NAKASUKA: Towards a Multiagent Design Principle -Analyzing an Organizational-Learning Oriented Classifer System-, Loia,V., Sessa,9.(Eds.) *Soft Computing Agents: New Trends for Designing Autonomous Systems*, Special Issue of the Series "Studies in Fuzziness and Soft Computing", Physica-Verlag, Springer Publisher, 2001

[18] S. Kurahashi and T. Terano: Analyzing Norm Emergence in Communal Sharing via Agent-Based Simulation, LNAI 2253, pp. 88-98 (2001).

[19] H. Sato, H. Matsui, I. Ono, H. Kita, T. Terano, H. Deguchi, and Y. Shiozawa: U-Mart Project: Learning Economic Principles from the Bottom by Both Human and Software Agents, LNAI 2253, pp. 121-131 (2001).

[20] H. Deguchi, T. Terano, K. Kurumatani, T. Yuzawa, S. Hashimoto, H. Matsui, A. Sashima, and T. Kaneda: Virtual Economy Simulation and Gaming -An Agent Based Approach-, LNAI 2253, pp. 218-226 (2001).

[21] A.Namatame, T. Terano, K. Kurumatani (eds.): *Agent-Based Approaches in Economic and Social Complex Systems*, IOS Press and Ohmsha (2002).

[22] S. Kurahashi, T. Terano: Emergence,Maintenance,and Collapse of Norms on Information Communal Sharing: Analysis via Agent-Based Simulation, *Agent-Based Approaches in Economic and Social Complex Systems*, IOS Press and Ohmsha, pp. 25-34 (2002).

[23] H. Sato, H. Matsui, I. Ono, H. Kita, T. Terano, H. Deguchi, Y.Shiozawa: Case Report on U-Mart Experimental System: Competition of Software Agents and Gaming Simulation with Human Agents, *Agent-Based Approaches in Economic and Social Complex Systems*, IOS Press and Ohmsha, pp .167-178 (2002).

[24] H. Deguchi, T. Terano, K.Kurumatani, T. Kaneda, H. Matsui, T. Yuzawa, A. Sashima, Y. Koyama, H. Lee, M. Kobayashi: Virtual Economy -Agent Based Modeling and Simulation of National Economy-, *Agent-Based Approaches in Economic and Social Complex Systems*, IOS Press and Ohmsha, pp198-207 (2002).

3.5 解説

[1] S. Yamada, H. Kawano: Information Gathering and Searching Approaches in the Web, *New Generation Computing*, Vol. 19, No. 2, pp. 195-208 (2001).

[2] 山田誠二，村田剛志，北村泰彦：知的 Web 情報システム，人工知能学会誌，Vol. 16, No. 4，pp. 495-502 (2001).

[3] 大澤幸生, 高間康史：WWW に潜む創造的意思決定のチャンス, 人工知能学会誌, Vol. 16, No. 4, pp. 530-534 (2001).

[4] 沼尾正行, 宮下和雄 (インタビュア): 社会へのアンテナ第 4 回 高橋邦芳 氏 – ユーザを納得させられる人工知能技術の実現を, 人工知能学会誌, Vol. 16, No. 5, pp. 732-733 (2001).

[5] 加藤直樹, 羽室行信, 矢田勝俊: 新規顧客からのロイヤルカスタマーの早期発見, ES-TRELA, No. 89, pp. 10-17 (2001).

[6] 大澤幸生: 「予兆発見の科学」, (株) 博報堂「未来予兆」(2002)

[7] 大澤幸生・松尾豊・松村真宏: 予兆発見 - 自然・社会からの意思決定, システム制御情報学会誌 (2002)

A01 班：アクティブ情報収集

対話的文書検索によるアクティブ情報収集

研究代表者　山田 誠二　(東京工業大学大学院総合理工学研究科)
研究協力者　岡部 正幸　(科学技術振興事業団)

はじめに

アクティブマイニングがアクティブとなるための要である**アクティブ情報収集**は，不特定・非定常・大規模・分散知識源の中から，ユーザの目的や興味に合致するデータやそれらの関連を効率良く探索し前処理するための情報収集を目指す[1]．このようなアクティブ情報収集のためには，ユーザの目的や興味に合致した質の高い情報をいかに効率よくたくさん集めるかが最重要課題の一つとなる．

我々は，この課題に対し，ユーザとのインタラクションを持ちながら，その興味にあった情報検索システムの開発が，アクティブ情報収集の実現のための最も重要な要素技術の一つであると考える．そこで，適合フィードバックと関係学習を用いて対話的に文書を検索するシステムの実現により，アクティブ情報集の実現のための基盤をつくる．本稿では，特に対話的文書検索を用いて，Web ページのフィルタリングを行う研究について報告する．

Web ページのフィルタリング

WWW の急速な普及によりインターネット上では日々多様な情報発信が行われている．検索エンジンは，これらネット上に散在する膨大な量の情報へのアクセスを可能としており，WWW を情報源として活用する上で欠かせないツールとなっている．検索エンジンは通常，ユーザから与えられる検索条件を用いて対象ページを絞り込み，それらをランキングしたものを**ヒットリスト**として返す．しかし，ユーザが検索エンジンに入力する単語は一般的に平均 2〜3 語と少なく [3]，多くの Web ページがヒットしてしまうため，それらを全て調べることは難しい．また，ヒットリストの上位にユーザの要求を満たす Web ページ（**適合ページ**と呼ぶ）が集中しているとは限らず，順位が低くても，適合ページがたくさん見つかる場合も多い．効率的な検索を行うには，ユーザの検索意図を反映したランキングを行うことが必要であるが，ランキング方法は，それぞれの検索エンジンが独自に行っており，ほとんどの場合，その設定をユーザが調節することはできない．そこで，本研究では，検索エンジンの返すヒットリストから適合ページのみを自動的に選別する処理である**フィルタリング**を行うことによって検索を効率化する．

しかし，一般にユーザは検索を始める際に，目的とする情報を得られる Web ページがどのような特徴を持っているかを明確には知らない．また，ある Web ページが目的に沿うものであるかどうか，つまり適合ページであるか否かを判断することができても，その理由を判別条件として提示することは負担のかかる作業であり，ユーザが適切な条件設定（**フィルタ生成**）を行うことは難しい．本研究では，このフィルタ生成をユーザにできるだけ負担をかけることなく行うために，**適合フィードバック** [11] と**関係学習**を用いた対話型の検索システムを提案する [7]．適合フィードバックは，文書検索の分野で提案された，

[1] http://www.ar.sanken.osaka-u.ac.jp/activemining/

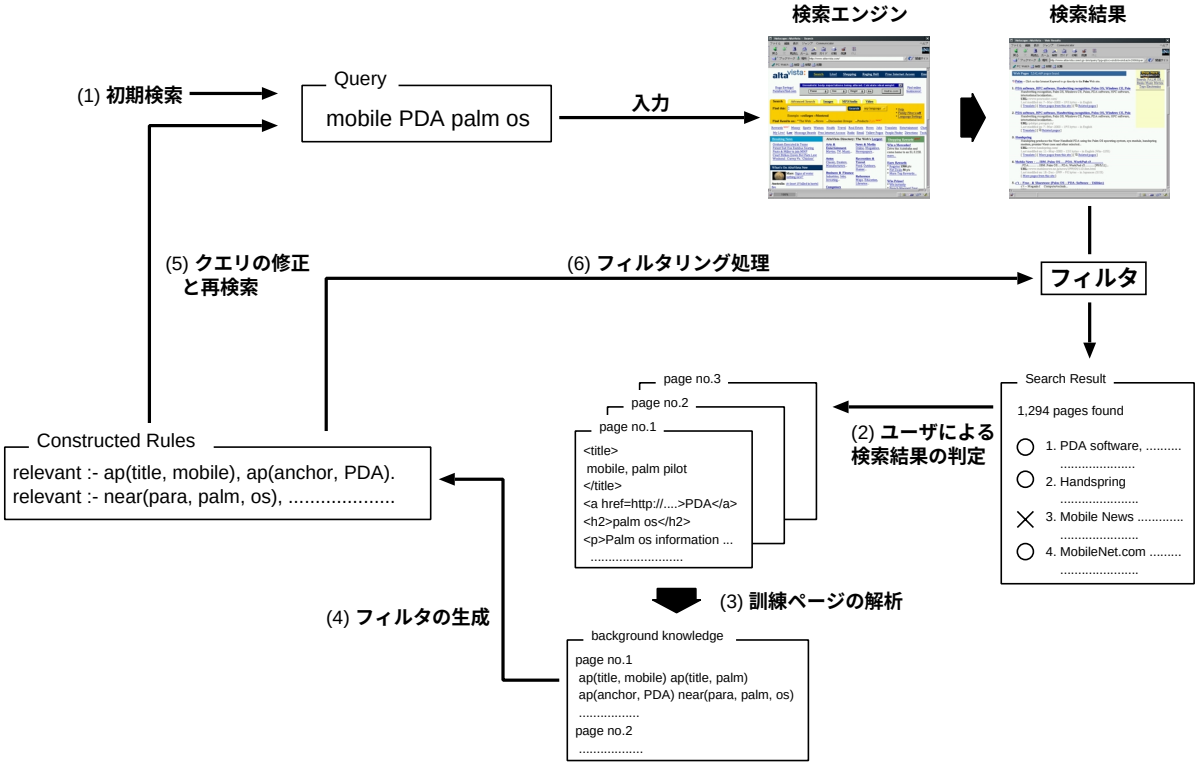


図 1: フィルタリングを伴う検索処理

ユーザの検索要求を自動的に推定するための枠組みで，ユーザによる文書判定とその情報を利用した再検索を繰り返しながら，徐々に適合文書を集めていくという対話的アプローチを提供する．この方法を用いることで，ユーザが適合文書の判定さえすれば，適宜判別条件を自動修正していくことが可能となる．

我々はこれまで，再検索時に新たな適合文書を獲得するために有効な単語間の関係を学習する方法を提案し，新聞記事を使った検索実験を通して，その効果を確認している [5, 6]．本研究では，この方法を拡張し，Web 検索エンジンが返すヒットリストから，適合ページのみを選びだすフィルタの自動生成へ適用する．なお，このフィルタは複数のルール集合で構成され，各ルールは，キーワード，論理演算子，近接演算子，タグ情報の組み合わせによって表現される．一般に，Web ページはタグ付けがなされており，タグの種類によってページ内のテキストの重要度が異なる [13]．よって，検索範囲としてタグを指定することで，Web ページの特徴を活用することができ，より効率的な絞り込みが行えると考えられる．

このように，教師付き学習を行うことによって，より精度の高い Web ページ収集を行うことを目的としたシステムは，これまでにいくつか提案されている．Syskill & Webert [9] は，情報利得を使ったキーワード抽出とベイズ分類器によるフィルタリング機能を持った Web ページ収集システムである．現在見ているページが持つリンクページの中から，もしくは検索エンジンに直接クエリを与えることにより，ユーザが興味を持つと思われるも

のを推薦する．このシステムはキーワードのみを特徴として用いており，本研究のようにキーワード間の関係やタグ情報などは考慮していない．また，検索エンジンが返すヒットリストのフィルタリングは行わず，そのまま用いている．Focused Crawler [1] は，フィルタリング機能を持ったWebページ収集ロボットである．このシステムは，通常の検索ロボットのように任意のページを収集するのではなく，特定の検索要求に見合うWebページのみを選択的に収集する．辿るべきリンクを決定するために，ユーザより与えられる訓練例を使ったベイズ分類学習を行う．学習はページ内に現れる単語の頻度に基づいて行われるため，キーワード間の関係やタグ情報は考慮されない．また，予め与えられる分類階層を利用しているため，本研究のように個別の検索要求に焦点を当てた検索には向かないシステムといえる．

以下の章では，まず2章でこのシステムを使った検索過程について説明する．次に3章でシステムの中心的な機能となるWebページの構造を利用したルールの表現と生成方法について述べ，4章で検索実験を行いシステムの能力を調べる．5章ではシステムの効果をより詳しく考察し，最後に6章で本研究をまとめる．

適合フィードバックによる対話的Web検索

図1は，本研究で提案するシステムを使った検索処理の概要である．以下，各ステップで行われる手続きについて述べる．各手続きは，図中の番号の付いた矢印における処理と対応しており，ユーザ側で行う操作とシステム側で行う操作の両方を記述している．

1. **初期検索**　初期条件として，検索エンジンに与える単語集合（以下**クエリ**と呼ぶ）と言語設定，日付指定等の入力をユーザに促し，入力された情報を検索エンジンに与え，検索結果を得る．

2. **ユーザによる検索結果の判定**　検索結果で上位にランクされたWebページ（通常上位10ページ程）をユーザに判定してもらい，適合ページ（**正例ページ**）と非適合ページ（**負例ページ**）に分けて，**訓練ページ**として保存する．

3. **訓練ページの解析**　フィルタを生成する際に必要な情報を訓練ページの解析により得る．具体的には，各キーワードのページ中における出現場所（タイトルやアンカーテキストなど）と近接しているキーワードの組み合せを各訓練ページ毎に調べ，リテラルを生成して，フィルタを構成する条件候補集合を作る．

4. **フィルタの生成**　(3)で得られた条件候補集合を使って，関係学習を行い，正例ページを含み負例ページを排除するフィルタを生成する．

5. **クエリの修正と再検索**　ページの判定や解析を行う中で，クエリに付け加えるべき単語などが見つかった場合，また適合ページが全く見つからない場合などに，クエリを修正して検索エンジンに与え，新たな検索結果を受け取る．

6. **フィルタリング処理**　検索結果で上位にランクされたWebページからフィルタリングを行い，フィルタを通過したページが必要数集まった時点で，その結果をユーザに提示する．ただし，既にユーザによって判定が行われたページは除く．

検索は以上の手順で進み，(6) から (2) へ戻ることによりフィードバックが繰り返される．フィードバックを更に繰り返すかどうかは，そのときの検索結果を評価するユーザ側の判断で行うことができ，最終的に十分な情報が得られれば検索は終了となる．

以上の手続きの内，提示されるページの順位に直接影響する操作は，(5) と (6) である．検索エンジンでは，これら 2 つの操作を支援するための機能を提供している場合が多い．(5) のクエリの修正については，関連単語を選ぶための方法が，情報検索の分野においてこれまでに数多く研究されている [8, 12]．検索エンジンの中には，MetaCrawler[2]のように実際にクエリの単語に関連した語をいくつか提示してくれるものもある．(6) のフィルタリングについては，オプションで複雑な論理式を指定することができるものなどはあるが，どのような設定をしたらいいのかを支援してくれる機能を提供するものは，今のところ存在しない．次章では，この (6) の操作を行うための Web ページの特徴を生かしたフィルタの表現と生成方法について詳しく述べる．

フィルタの表現と生成アルゴリズム

フィルタは複数のルールから構成され，各ルールはユーザから提示された正例ページと負例ページを訓練例とする分類学習を行うことにより得られる．本章では，まずルールの表現形式について述べ，次にその生成方法を示す．

ルールの表現

学習により獲得するルールは，キーワード，演算子，検索範囲の指定がなされたホーン節で表現する．ルールの条件部を構成するリテラルには，次のものを用いる．

- $ap(region\_type, word)$： ページ内の $region\_type$ 部分に $word$ が現れる．
- $near(region\_type, word1, word2)$： ページ内の $region\_type$ 部分で $word1$ と $word2$ が 10 単語以内に順不同で近接して現れる．

演算子は，単語間の基本的な位置関係を表現する．近接関係は以前からその有効性が確認されており [4]，近年この関係を指定，または自動的に考慮する検索エンジンが増えている．また，Web 検索では，同じ単語でもページ内における出現場所によってその重要度が異なると考えられる．例えば，タイトルタグ内のテキストはそのページの主題を表現していることが多く，重要な手かがりとなる．よって ap，$near$ リテラルともに $region\_type$ を加えることによって，より詳しい位置関係を指定している．$region\_type$ の種類は，以下のものである．

- $title$： `<TITLE>`タグで囲まれたテキスト．
- $anchor$： `<A>`タグで囲まれたテキスト．
- $heading$： `<H`n`>`（$n = 1$〜4）タグで囲まれたテキスト．

[2] http://www.metacrawler.com

- $para$: <P>タグで囲まれた 20 語以上からなるテキスト.

これらのリテラルにより，例えば次のようなルール集合が生成される.

$$\begin{cases} relevant \text{ :- } ap(title, \text{mobile}), ap(anchor, \text{PDA}). \\ relevant \text{ :- } near(para, \text{palm}, \text{os}). \end{cases}$$

各ルールは OR 関係にあり，複数のルールの内一つでも満たせば適合ページと判定する．上のルール集合は，ページのタイトルに “mobile” が現れ，かつページ内に “PDA” が現れるアンカーテキストが存在するページ，またはページ内の同一段落で “palm” と “os” が近接して現れているページを表している.

ルール集合の生成

フィルタとなるルール集合 R を生成するための手続きを図 2 に示す.

Separate-and-Conquer 戦略

この手続きは，Separate-and-Conquer 戦略 [2] を用いており，ルール（図 2 の $rule$）を一つずつ生成し，R に追加する作業を繰り返す．$rule$ が一つ生成されると，それによって被覆される文書が正例文書集合 E^+ から取り除かれるので，$rule$ が生成される度に E^+ は減少していき，最終的に空集合となれば手続きが終了となる．同じ正例ページであっても文書中で使われる単語や，近接して現れる単語の組合わせが違うこともあり，そのページを識別するために有効な特徴は正例ページによって異なる.

重み付き情報利得を用いたルール生成

各ルールは空のボディ部にリテラルを一つずつ追加していき，負例を一つも含まなくなると完成となる．追加するリテラルは，条件候補リテラル集合 C の中から選ばれる．ここで C は，キーワード集合 K と $region\_type$ を引数に代入することにより作られる全てのリテラルの内，訓練ページで実際に成り立つものの集合を指す．具体的には次のようなリテラルである.

- K のすべての要素を引数 $word$ に代入した ap リテラルを各 $region\_type$ ごとに生成したものの中で，少なくとも 1 つの正例ページで成り立つもの.
- K の要素のすべてのペアを引数 $word1$, $word2$ に代入した $near$ リテラルを各 $region\_type$ ごとに生成したものの中で，少なくとも 1 つの正例ページで成り立つもの.

また，追加するリテラルを選択する際の評価基準には，以下の式から計算される重み付き情報利得 G を用いる [10].

$$G = e^{\oplus}_{new}\{I(e^{\oplus}_{old}, e^{\ominus}_{old}) - I(e^{\oplus}_{new}, e^{\ominus}_{new})\}$$

$$I(e^{\oplus}, e^{\ominus}) = -log_2 \frac{e^{\oplus}}{e^{\oplus} + e^{\ominus}}$$

$e^{\oplus}_{old}$, $e^{\ominus}_{old}$, $e^{\oplus}_{new}$, $e^{\ominus}_{new}$ はそれぞれ，リテラル追加前と追加後に満たす正例ページと負例ページの数である．これにより，正例ページ 1 つあたりの情報利得が大きく，かつ正例ページをより多く満たすリテラルが選ばれ，追加される．なお，G が最大となるリテラルが複数存在する場合，ランダムに選択する.

入力：正例ページ集合 E^+，負例ページ集合 E^-，
　　条件候補リテラル集合 C，キーワード集合 K.
出力：ルール集合 R
変数：ルール $rule$,，除外リテラル l_1,
　　除外リテラル集合 S.
初期化：
　$K \leftarrow$ 検索式内の単語集合
　$R,\ S,\ l_1 \leftarrow empty$
　$rule \ \leftarrow \ relevant$:-
Repeat
　· $rule$ を満たす正例ページ数 p と負例ページ数 n を調べる.
　if $n = 0$ **then**
　　· $rule$ を R に加える.
　　· $rule$ を満たす正例を E^+ から取り除く.
　　if E^+が空集合 **then 終了**
　　else $rule$，S，l_1 を初期化.
　else
　　· S 中のリテラルを除く C 中の全てのリテラルについて,
　　重み付け情報利得 G を計算する.
　　if $G > 0$ となるリテラルがない **then**
　　　if $rule$ のボディ部が空 **then**
　　　　· K にキーワードを一つ加える.
　　　　· C を新しく生成する.
　　　else
　　　　· S と $rule$ を初期化する.
　　　　· l_1 を S に加え，l_1 を初期化.
　　else
　　　· G が最大となるリテラルを l_{max} とする.
　　　if $rule$ のボディ部が空 **then** $l_1 := l_{max}$
　　　· l_{max} を $rule$ と S に加える.

図 2: ルール集合生成手続き

バックトラックとキーワードの追加

ルール生成途中では，全ての正例を満たすルール集合が生成されないまま，リテラルの選択候補がなくなり，探索が止まることがある．この時，$rule$ のボディ部にリテラルが 1 つ以上追加された状態であれば，その時点で $rule$ に追加されているリテラルを全て破棄し，$rule$ の生成をやり直す．その際，同じ探索を繰り返さないため，生成をやり直す前の

<num> Number: 401
<title> foreign minorities, Germany

<desc> Description:
What language and cultural differences impede the integration of foreign minorities in Germany?

<narr> Narrative:
A relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of immigration difficulties is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant.

図 3: トピックの例

$rule$ に追加されたリテラルのうち，最初に追加されたもの（図 2 中の変数 l_1）を予め除外しておく．そうでない場合，つまり $rule$ のボディ部が空の状態である場合，K にキーワードを新たに加え，C を新しく生成することによって選択可能なリテラルを作る．

追加するキーワードは，正例ページ集合 E^+ から選ぶ．まず各ページ中の<P> タグで囲まれた 20 単語以上からなる段落の内，クエリ内の単語を少なくとも 1 つ含むもののみを集め，これを T とする．次に T に出現する全単語集合 W の各要素 w_i に対して，以下の式から重要度を計算する．

（w_i の重要度）＝（T 中における w_i 平均出現頻度）×（w_i が現れる T 中の段落数）

この重要度が最も高いもので，クエリに使われておらず，まだ追加されてないものを新しく追加する．

実験

提案システムの有効性を調べるために，2 章で説明した手順に従った検索実験を行った．

実験方法

検索エンジンには，検索精度が良いとされる Google[3] を用い，英語で記述されたページを検索対象とした．また，コンテスト形式による検索システムの性能評価を目的とした会議である TREC[4]の Small Web Track において利用された検索課題（トピック）から 20 個 (No.401〜420) を選んで用いた．

図 3 は実験で用いたトピックの例で，検索要求や適合文書とする際の判定基準などが記述されている．適合ページの判定は，これらの記述に従い全て同一人物が行った．各トピックの<title>タグには，1〜3 単語が記されており，これを初期クエリとして与えた．また，この実験は，システムが生成するフィルタの性能評価を目的とするため，2 章で述べた手続き 5 の処理は行わない．

[3] http://www.google.com
[4] http://trec.nist.gov

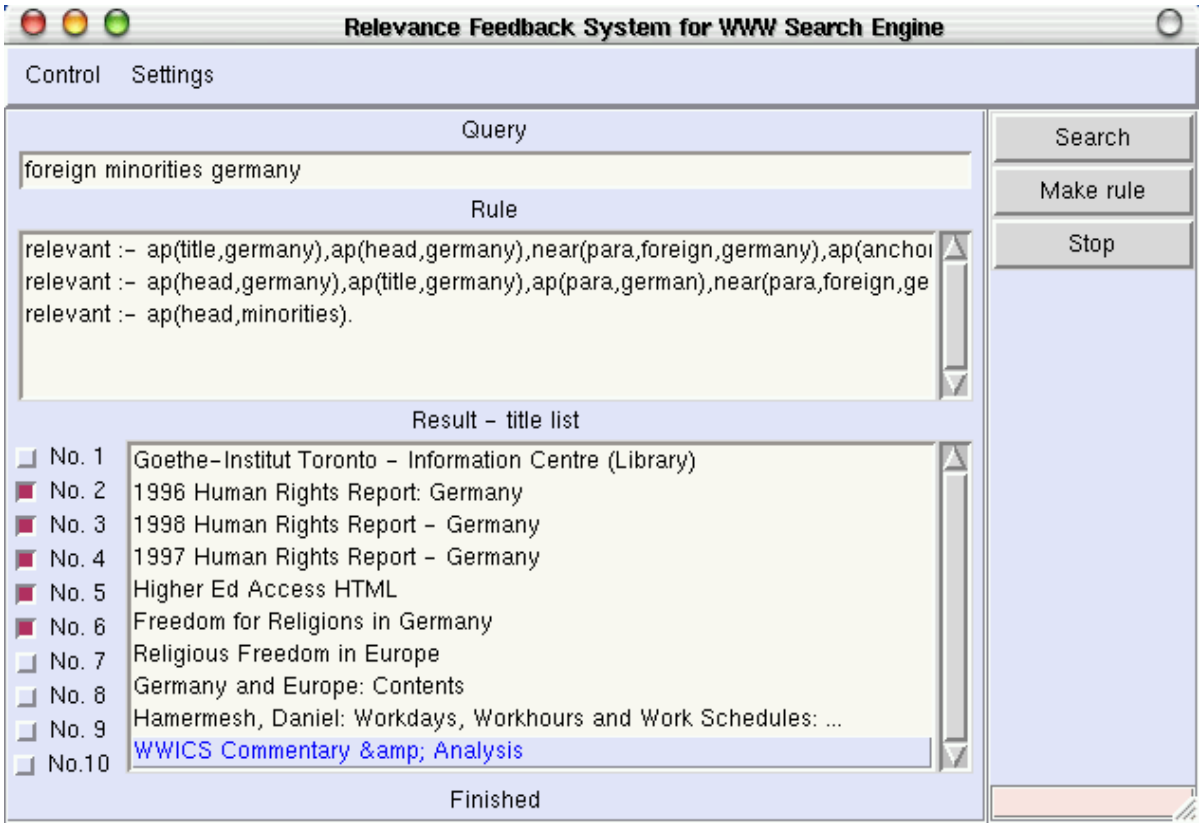


図 4: システムインタフェース

評価方法として，Google のみを使った通常の検索（フィルタなし Web 検索）とフィルタを用いた提案システムによる検索（フィルタあり Web 検索）について，一定数のページを見た場合に得られた適合ページ数をトピック別に比較した．どちらの検索も，各トピックにつき合計 50 ページを調べた．Google のみによる検索では，初期検索結果の上位 50 ページを調べた．提案システムを用いた検索では，10 ページ判定する毎にフィルタを新しく生成し，ヒットリストの上位からまだ調べていないページについてフィルタリングを行い，10 ページ得られた時点で判定を行う操作を 4 回繰り返した．

図 4 は，提案システムのインタフェースである．“Query” 部分に入力された単語が検索エンジンに入力され，返ってきたヒットリストをフィルタリングしたものの上位 10 ページのタイトルが “Result” 部分に表示される．タイトルをクリックすると，ブラウザが立ち上がりそのページが閲覧できるようになっている．ページをみた後，タイトル番号横のボタンを押すことでそのページが適合ページであることをマークする．マークがないものは全て非適合ページとして扱われる．このあと “Make Rule” ボタンを押すことでフィルタルールが生成され，“Rule” 部分に表示される．

実験結果

図 5 は，フィルタあり Web 検索とフィルタなし Web 検索を行った際の，判定ページ数と獲得適合ページ数の関係を示したものである．各値は，20 個のトピックについての平

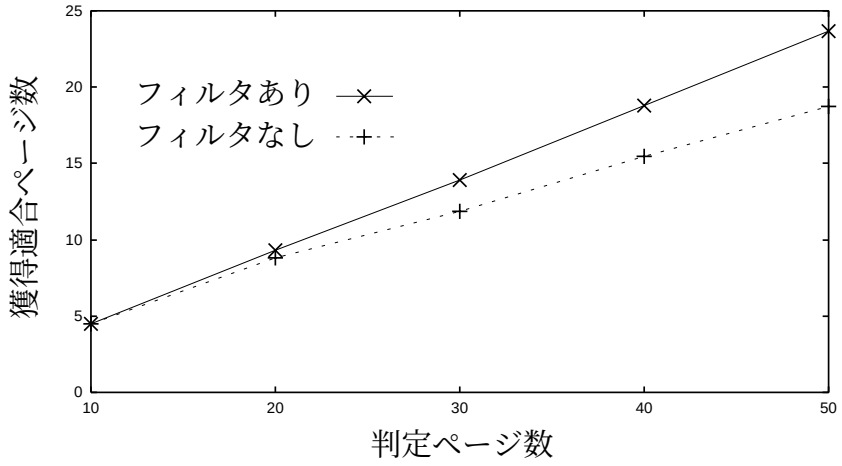


図 5: 獲得適合ページの平均値

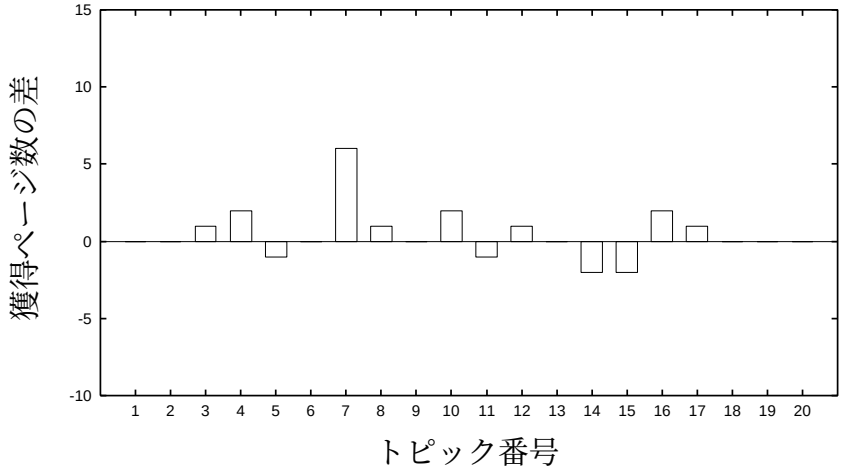


図 6: トピック別獲得適合ページ数の差（20 ページ判定後）

均値である．最初の 10 ページはどちらも同じものを評価するので差はないが，それ以降提案システムを用いたものは，フィルタリングによって提示されるページが違うため，差が生じている．50 ページを判定した段階でフィルタあり Web 検索は平均 5 ページ多く適合ページが得られている．このように平均的に見てフィルタリングの効果が現れていると言えるが，トピックによってその効果は大きく異なる．

図 6〜図 9 は獲得適合ページ数の差 D をトピック別に示したものである．フィルタあり Web 検索により得られた適合ページ数を A，フィルタなしの場合の適合ページ数を B とすると，$D = A - B$ で示される．図 6 では，判定ページ数が少ないので正例ページをあまり得られないトピックが多くあること，また検索エンジンのヒット率もまだ高いこと等が原因でフィルタリングの効果はほとんど見られず，逆効果が現れているトピックも 4 つある．図 7 では多くのトピックで効果が現れ始めている．さらに図 8 では，図 7 で効果

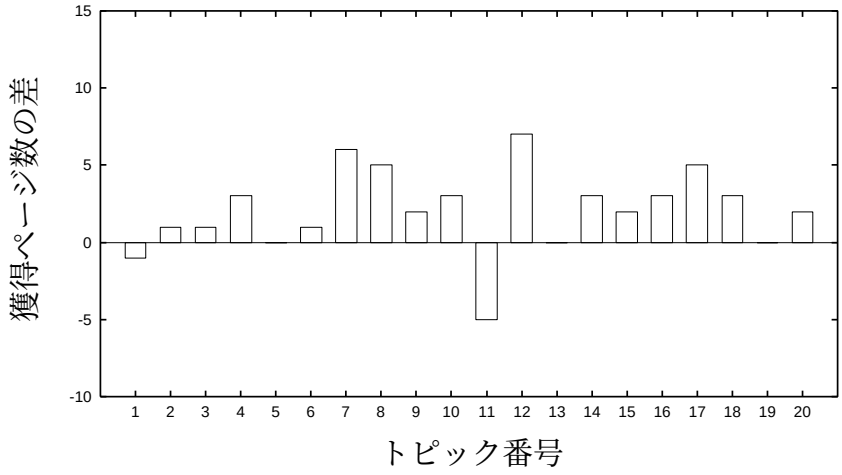


図 7: トピック別獲得適合ページ数の差（30 ページ判定後）

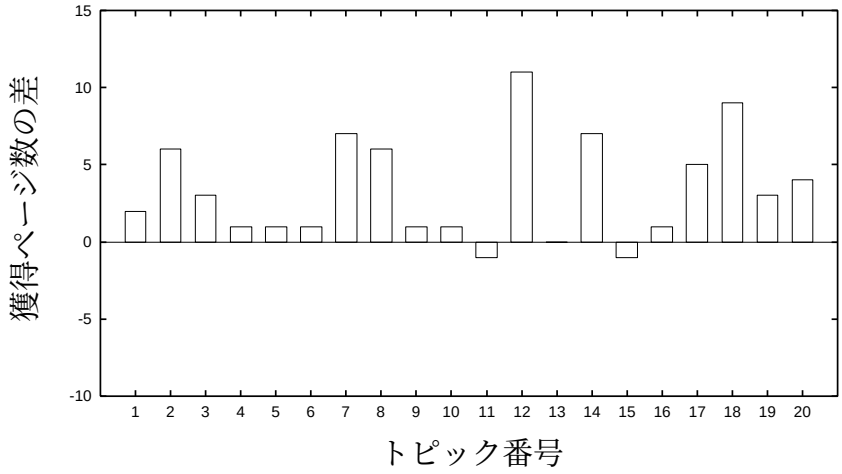


図 8: トピック別獲得適合ページ数の差（40 ページ判定後）

の出たトピックの多くで差が広がっている．図 9 では効果が得られているものとそうでないものとの差がはっきり分かるようになる．効果の現れ方はトピックによってまちまちであり，4，10，11，13，15 番のトピックのように差が現れないものや負の値となるものがあるものの，判定ページの増加と共にほとんどのトピックについて差が正の値となっており，全体的に見て，フィルタリングの効果が現れているといえる．

考察

実験結果から，ほとんどのトピックにおいて，獲得適合ページ数の差が正の値になっているが，約半数のトピックについては，検索エンジンのみでも十分な適合ページが得られるため目立った効果が現れなかった．ここでは，うまく効果が現れたものとそうでないものの例を示す．

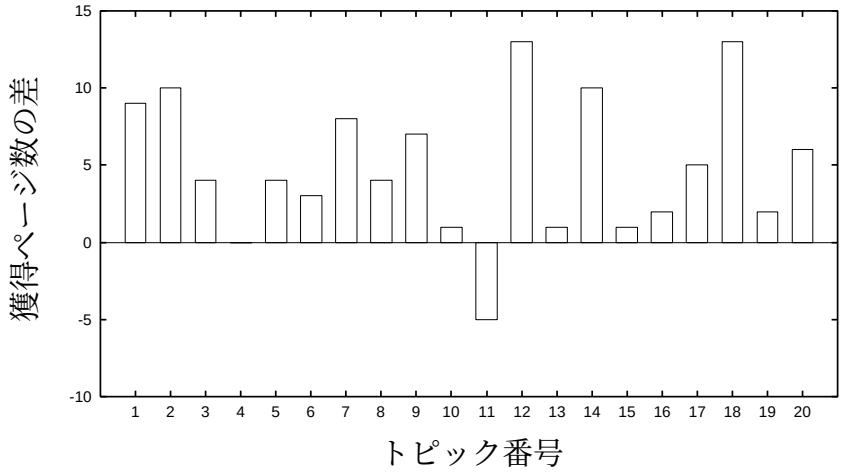


図 9: トピック別獲得適合ページ数の差（50 ページ判定後）

12 番と 18 番のトピックは，フィルタリングの効果が一番良く現れている．12 番の検索要求は,「空港における有効なセキュリティシステムにはどのようなものがあるか」というものである．検索エンジンがこの問に関して返すヒットリストのなかには，適合ページの他に,「旅行者が心得ておくべき注意点」を紹介しているページが多く含まれていた．フィルタリングを使った検索では，これらの非適合ページを多く排除できたため効果が大きく現れた．表 1 は，このトピックに関する検索で生成されたルールの内，適合ページを多く獲得したものの一部である（各ルールともボディ部のみを示してある）．各ルールをみてわかるように，“airport” と “security” を基本に，“faa” や “screening” といった具体的なシステムを意味する単語が組み合わされることにより，精度の高いフィルタリングが行われていた．

また，18 番の検索要求は,「キルト製品でどのようにして収入が得られているか」というもので，キルトについて書かれた本，キルト教室等を紹介したものが適合ページとなる．フィルタリングを使った検索では，これらのページの特徴をうまく捕らえて効果を上げており，特にキルト製品をオンライン販売しているページが多く検索されていた．表 2 に，同じく適合ページを多く獲得したルールの一部を示す．この中で，“online” と “quilt” の 2 つの単語が用いられているルールにより，オンラインショップ関連のページが選び出されていた．また，“fabric” と “quilt” が用いられているルールで得られたページでは，織物コレクションの一部としてキルトが紹介されていた．

逆に効果が全く見られなかった 11 番のトピックは,「難破船を引き上げ宝物を得るために必要な情報」が得られるページを探すというものである．このトピックに適合するページは，リンク集，掲示板，ニュース，宝探しのグループのホームページなどの多種類のページが少しずつあるため，共通する特徴を見つけるのが困難であり，個々のページに特化された特徴が生成されてしまうことから，新しい適合ページを得ることができず，効果が出なかったと考えられる．表 3 には，生成されたルールの中で適合ページが得られたものを示してある．これらを見てわかるように，生成されたルール数が少なく，条件部も

表 1: トピック 12 番で生成されたルールの例

```
:- ap(anchor,screening).
:- near(para,security,system),ap(title,airport).
:- near(para,security,airports),
   near(para,security,access).
:- near(para,security,airports),
   near(para,faa,system).
```

表 2: トピック 18 番で生成されたルールの例

```
   :- ap(para,online),ap(title,quilts).
   :- ap(anchor,online),ap(title,quilting),
      ap(anchor,quilting).
   :- ap(para,block),near(para,quilt,block),
      ap(anchor,fabric).
   :- ap(title,quilting),ap(anchor,fabric).
```

表 3: トピック 11 番で生成されたルールの例

```
:- ap(anchor,shipwreck).
:- ap(anchor,shipwreck),ap(anchor,salvaging).
```

Web ページを絞り込むには不十分であるため，フィルタリングの効果が現れなかった．

まとめ

本研究では，アクティブ情報収集実現の要素技術として，関係学習による対話的文書検索を提案した．検索エンジンが返すヒットリストを逐次的にフィルタリングすることによって，適合ページを効率よく選別する対話的な検索処理システムの構成とその検索手続きについて説明した．提案したシステムは，Web ページ中のキーワードの位置関係と構造的条件を加味してフィルタを自動生成し，複雑な絞り込みを行うことができる．本研究では，これを実験を通して確認することができた．

現行の検索エンジンでは，このようにユーザからのフィードバック情報を処理し，ユーザ個別の情報検索を支援する枠組みはまだ提供されておらず，検索エンジンをより有効に活用するために本研究で述べたアプローチは十分有効であると考えられる．

今後の課題としては，ユーザがページの判定をスムーズに行うためのインタフェースの工夫や視覚化機能を追加することである．また，適合フィードバックを行う際にユーザの負担となる判定ページの必要数をできるだけ減らすことは重要であり，クラスタリング機能などの追加を検討している．

参考文献

[1] S. Chakrabarti, M. Berg and B, Dom, "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery," Proc 8th Int. World Wide Web Conf, 1999.

[2] J. Furnkranz, "Separate-and-Conquer Rule Learning," Artificial Intelligence Review, Vol.13, No.1, 1999.

[3] M.B. Jansen, A. Spink, J. Bateman and T. Saracevic "Real life information retrieval: A study of user queries on the web," SIGIR Forum, Vol.32, No.1, pp.5-17, 1998.

[4] E.M. Keen, "Some aspects of proximity searching in text retrieval system", Journal of Information Science, Vol.18, No.2, pp.89-98, 1992.

[5] 岡部正幸，山田誠二，"関係学習を用いた対話的文書検索，" 人工知能学会誌，Vol.16, No.1P, 2001.

[6] M. Okabe and S. Yamada, "Interactive Document Retrieval with Relational Learning," Proc. 16th ACM Symposium on Applied Computing, 2001.

[7] M. Okabe and S. Yamada: Interactive Web Page Filtering with Relational Learning, The First Asia-Pacific Conference on Web Intelligence (WI-2001), pp.443–447 (2001).

[8] x1 M. Mitra, A. Singhal and C. Buckley, "Improving automatic query expansion," Proc. 21st annual international ACM SIGIR, pp.206-214, 1998.

[9] M. Pazzani, J. Muramatsu and D. Billsus, "Syskill & Webert: Identifying interesting web sites," Proc. AAAI, 1996.

[10] J.R. Quinlan and R.M. Cameron-Jones, "Induction of Logic Programs: FOIL and Related Systems," New Generation Computing, Vol.13, Nos.3,4, pp.287-312, 1995.

[11] G. Salton and C. Buckley, "Improving Retrieval Performance by Relevance Feedback," Journal of the American Society for Information Science, Vol.41, No.4, pp.288-297, 1990.

[12] J. Xu and W.B. Croft, "Query expansion using local and global document analysis," Proc. 19th annual international ACM SIGIR, pp.4-11, 1996.

[13] D. Zhang and D. Yisheng, "An efficient algorithm to rank Web resources", Proc 9th Int. World Wide Web Conf, pp.449-455, 2000.

WWW 上の情報収集／可視化のための免疫ネットワークを用いたクラスタリング

研究分担者　高間 康史　(東京工業大学大学院総合理工学研究科)
研究協力者　廣田 薫　(東京工業大学大学院総合理工学研究科)

背景と目的

WWW 上で多数公開されているオンラインニュース記事や，ユーザによる一連の情報検索作業により得られる検索結果など，WWW 上の情報収集においては時系列的関連を持った文書集合族を対象とすることが多い点に着目し，これを扱うための可塑的クラスタリング手法について提案し，情報収集，情報可視化システムへの応用について検討する．

膨大かつ多様な情報が存在する様になった現在の WWW 空間においては，ブラウザやサーチエンジンなどの普及により，情報入手にかかるコストは非常に低下しており，ユーザの利用形態も多様になりつつある．しかし，従来型検索エンジンのほとんどにおいて，検索結果はユーザの入力したクエリーとの（何らかの特徴・尺度を用いて計算した）類似度に基づいて線形にランク付けしたリスト形式で返されるが，クエリーにマッチする情報が多数存在する場合のユーザ負荷が膨大であることが従来より指摘されている．この問題に対し，検索結果をクラスタリングして提示したり [3, 16, 17]，関連する話題分布を空間配置したキーワード（キーワードマップ）[9] を用いて提示する事により，ユーザの情報検索作業を支援する**情報可視化システム** [3, 9, 11, 12, 16, 17] の有効性が指摘され，研究が進められている．

しかし，クラスタリングを用いた情報可視化システムに関する従来研究では，ディレクトリサービス（Yahoo!カテゴリなど）の様に静的なカテゴリ構造を用いるか，あるいは検索結果毎にクラスタリング手法を独立に適用するかのどちらかであり，いずれも，一連の情報検索作業を通じたユーザのコンテキストを考慮していないという問題点がある．すなわち，ユーザの行う情報検索作業は，１回毎に独立したものでも，あるいは（適合フィードバックなどで仮定されるように）首尾一貫した検索要求の基に繰り返されるものでもない．検索結果を吟味し，それによって検索対象に関する知識や考えに変化が生じ，これに基づいて次のクエリーを決定する，という，サイクルを繰り返すことが多く [2]，この様な検索結果間の緩やかな時系列的関連性を効果的に捉える事のできるクラスタリング手法が必要である．

時系列的関連性を持った文書集合族という点では，WWW 上のオンラインニュースサイトで時々刻々と更新・配信されるオンラインニュース記事も同様であり，ある期間（１日，１週間など）に配信されたニュース記事集合の話題分布構造を，前後期間（前日・先週など）との時系列的な対応付けを考慮して可視化する事ができれば，話題の変遷や流行などを捉える上で有効である．

本研究では，この様な文書集合間の時系列的関連を考慮したクラスタリング手法（**可塑的クラスタリング**）について提案する．また，情報検索結果の可視化の際には，閲覧支援という面から文書のクラスタリングが重要視される傾向にあるが，オンラインニュース記事の場合には，キーワードマップの方が重要であろう．従来の情報可視化システムでは，

文書クラスタリングあるいはキーワードマップのいずれかに重点をおいたシステムが多いが，本研究で提案する手法は，文書クラスタリングを考慮しつつ，キーワード空間の可視化を行う点でも特徴的である．

提案システムでは，検索結果文書から抽出した名詞キーワードをネットワーク構造に漸進的に追加していき，現在の検索結果に対応したキーワードが活性化するように活性伝搬を行う．この際，同一文書に含まれるキーワード同士が同時に活性化しないような制約を加えることにより，文書クラスタリングも同時に考慮する．この様な活性伝搬を実現するためには，漸進的組織化，記憶保持，多様性の維持と言った特性が必要となるため，本研究では免疫ネットワークモデルを工学的に応用する．また，文書集合間の時系列的関連性については，免疫記憶細胞モデルを用いて考慮する．

検討内容

可塑的クラスタリング

本研究では，文書クラスタリングとキーワードマップにおけるランドマーク抽出は表裏一体の問題と考える．すなわち，同一キーワードを含む文書集合をクラスタ候補とし，文書全体のほとんどをカバーし，互いにオーバラップしないクラスタ集合を求めることでクラスタリングを行う．各クラスタに対応するキーワードがランドマークとなる．k-means などと異なり，クラスタ境界がキーワードの有無で厳密に決まるため，全ての文書が必ず一つのクラスタに属する事を保証できないが，異なる検索結果間の対応付けが容易であり，さらには WWW 上で実運用されているサーチエンジンのほとんどで採用されている，ブーリアンモデルとの相性がよいなどのメリットがある．また，同一話題に対応するキーワードを統一するため，既使用キーワードを優先して使用する事も重要である [11]. 提案手法は，異なる検索結果間のゆるい対応を考慮して柔軟なクラスタリングを行うことから，**可塑的クラスタリング**と呼ぶ．

互いにオーバラップしないクラスタ集合を抽出するだけであれば，組み合わせ最適化問題の一種として定式化でき，GA などにより準最適解を効率的に求めることが期待できる．しかし，各クラスタの持つ意味 (概念・話題) や，キーワードマップ上でのランドマークとしての観点からは，他の評価基準も要求される．すなわち，キーワードマップ上で近くに配置されている (出現文書集合が類似の) キーワード集合から，それぞれ代表となるものをランドマークとして抽出することが望ましい．本研究では，キーワード間の局所的相互作用に基づきキーワードの活性値を計算し，これに基づいてランドマーク (クラスタ中心) となるキーワード集合を抽出する．相互作用モデルとしては以下に示す免疫ネットワークモデルを採用する．

免疫ネットワークモデル

免疫ネットワークモデルは，免疫システムの多様性，記憶能力などを説明するために Jerne によって提案された仮説であり [4]，個々の抗体が互いに認識しあってネットワークを構成し，一度経験した抗原に対しては，抗原が消滅した後も対応する抗体およびそれを産生する B 細胞が，他抗体・B 細胞と刺激を与えあう事によって長く体内に留まる事ができ，結果として免疫の記憶，多様性が保たれるとしたものである．免疫ネットワーク

に関しては，抗原・抗体濃度変化を微分方程式で表す数理モデルなどが提案されており [1, 7, 10]，本研究では以下に示す比較的簡単なモデルを採用する．

$$\frac{dX_i}{dt} = s + X_i(f(h_i^b) - k_b) \tag{1}$$

$$h_i^b = \sum_j J_{ij}^b X_j + \sum_j J_{ij}^g A_j \tag{2}$$

$$\frac{dA_i}{dt} = (r - k_g h_i^g) X_i \tag{3}$$

$$h_i^g = \sum_j J_{ji}^g X_j \tag{4}$$

$$f(h) = p \frac{h}{(h+\theta_1)} \frac{\theta_2}{(h+\theta_2)} \tag{5}$$

ここで，X_i が抗体濃度，A_i は抗原濃度をそれぞれ表す (初期濃度 $X_i(0), A_i(0)$)．s は抗体の補充率，r は抗原の再生率，k_b, k_g はそれぞれ，抗体，抗原の死滅率である．h_i^b, h_i^g は field と呼ばれ，認識可能な抗原，抗体からの影響は式 (5) より，field の対数を横軸とするベル型の関数により定義される [11]．J_{ij}^b は，抗体 i, j 間の親和度，J_{ij}^g は抗体 i と抗原 j 間の親和度を表す．

式 (5) より，抗体は他抗体，抗原を認識して活性化するだけでなく，影響が強すぎる場合には抑制される．抗体活性に関するこの様な非線形により，自己由来の細胞は攻撃しない免疫寛容や，類似抗体間の競合による多様性の維持が説明できる．

免疫ネットワークモデルの安定性，ダイナミクスについて，ネットワーク構造やトポロジーを固定することにより解析が行われているが [1, 7, 10]，本研究では次節で示すように文書集合中のキーワード分布に基づいてネットワークを構築するため，上記の解析手法を適用する事は困難である．しかし，接続されたノード (抗体) 間で，安定してとりうる活性状態の組合わせについての考察により，以下の様な制約が示されている [12]．

- 抗体のとりうる活性状態は，初期状態，抑制状態，弱活性状態，高活性状態の 4 通りである．
- 接続された抗体がともに高活性化する状態は不安定である．
- 同一の高活性抗体と接続する抗体が複数存在する時，強接続の抗体は抑制され，弱接続の抗体は弱活性化する．

キーワードの活性度計算アルゴリズム

本研究では，キーワードを抗体，文書を抗原と見なすことにより，免疫ネットワークモデル (式 (1)–(5)) に基づいてキーワードの活性値を計算する．実際の免疫システムでは，侵入抗原に応じて抗体の構造 (認識対象) が変化するが，提案アルゴリズムでは文書集合に応じてリンク関係が変化することがこれに対応する．具体的な処理手順は以下の通りである．ここで，ネットワークの定常状態とは，高活性化するキーワード集合が一定となった状態を指すとする．

表 1: 実験に使用したパラメータ

Parameter	s	r	k_b	k_g	p
Value	10	0.01	0.4	10^{-4}	1.0
Parameter	T_k	T_d	$X_i(0)$	$A_i(0)$	TH_1
Value	3	3	10	10^5	3
Parameter	θ_1	θ_2	SC	WC	TH_2
Value	10^3	10^6	1.0	10^{-3}	3

1. 文書集合から，出現文書数 (document frequency) DF が TH_1 以上のキーワードを抽出し，ノードとする．出現文書集合が等しいキーワードについては一つにまとめる．
2. キーワード間接続強度 (親和度) (J_{ij}^b) の設定
   - 強接続 (SC)... 共起文書数 $CDF_{ij} \geq TH_2$
   - 弱接続 (WC)... $1 \leq CDF_{ij} < TH_2$
3. キーワード・文書間接続強度 (J_{ij}^g) の設定
   - 強接続 (SC)... 文書 j 中のキーワード i の頻度 $TF_{ij} \geq TH_2$
   - 弱接続 (WC)... $1 \leq TF_{ij} < TH_2$
4. キーワード，文書の活性値計算 X_i, A_i を式 (1–5) に基づき，ネットワークが定常状態になるまで繰り返す．

免疫記憶細胞のモデル化

オンラインニュース記事や，ユーザによる一連の情報検索作業により得られる検索結果などの，時系列的関連性を持った文書集合族から，話題の変遷や，ユーザの検索コンテキストをとらえるためには，以下の特徴を持つ話題を発見する事が重要となる．ここで，話題は文書中から抽出した名詞キーワードにより表現されると仮定する．

- 連続した複数の文書集合において含まれる話題（mainstream topic）
- 初期の文書集合に出現したものの一度消滅し，後に再び出現した話題（missing topic）

前者の話題は，情報検索であればユーザの興味に合致した，主要な話題と見なすことができる．mainstream topic は，未知の分野について検索を行っている場合には，対象分野の主要話題・概念を理解する上で重要なてがかりとなる．また，オンラインニュース記事の場合には，流行などに関連する事が期待できる．

また，missing topic の場合には，情報検索の初期段階でユーザが見落としてしまった話題である可能性がある．検索を続けていく内に再び出現するということは，検索対象において重要な話題であるかも知れないため，ユーザに再吟味を促す必要がある．また，オ

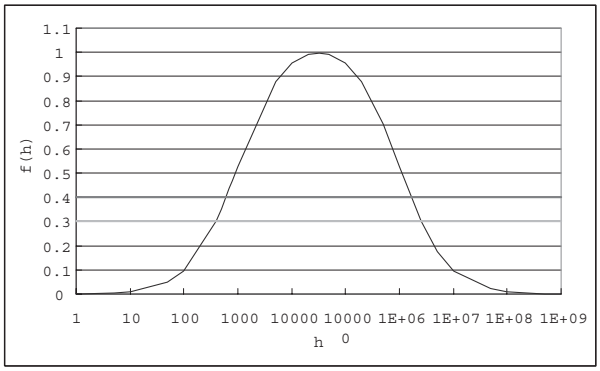


図 1: f(h) の形状と活性範囲の関係

ンラインニュース記事の場合には,(タイムスパンにもよるが) 流行の繰り返しなどに対応する可能性がある.

話題に関するこの様な特性を発見するために，提案手法では異なる文書集合において得られるクラスタ間の対応付けを行う．具体的には，以前に処理した文書集合において活性化したキーワード（クラスタ識別子）が優先して活性化するようにし，同様の話題が含まれる場合には同じクラスタ識別子が活性化するようにする事で，文書集合間のクラスタ対応付けを行う．これは，免疫ネットワークとのアナロジーにより，免疫記憶細胞を用いてモデル化可能である.

免疫記憶細胞の機能を考慮した数理モデルはいくつか存在するが，本稿では非常に簡単な手法を提案する [12, 13]．すなわち，免疫記憶細胞の死滅率を通常の抗体より低くするか，あるいは式 (5) のすその広がりを広くする（θ_1 を小さく，θ_2 を大きくする）ことで抗体の活性化範囲を広げることができるが，ここでは前者を採用する.

式 (1) より，式 (5) で表される関数 $y = f(x)$ （x は field）と直線 $y = k_b$ の交点より，抗体の活性範囲が決定されるが，表 1 のパラメータより通常の抗体の活性範囲は 607 から 1.65×10^6 となり，死滅率を $k_b = 0.3$ に下げると活性範囲は 395 から 2.53×10^6 に拡大する（図 1）．活性化範囲を広くした抗体は，通常抗体よりも優先して活性化することは，実験により確認されている [13].

結果

本節では、提案クラスタリング手法について，代表的なクラスタリング手法であり，有効性が実応用において広く認められている，k-means[5] と比較した結果について示す．ここでは，文書集合に独立に提案手法を適用した場合についての評価であり，時系列性は考慮していない.

k-means は対象データ（文書）全てをクラスタリングするのに対し，提案手法で生成するクラスタ構造は必ずしも全文書をカバーするとは限らない（経験的に $60 - 80\%$ の文書をカバーするクラスタ構造を生成 [11]）ため，この点で比較しても意味がない.

そこで本稿では，生成されたクラスタの結束性，理解容易性について，アンケート調

査した結果について示す．使用したデータは，Yahoo! Japan News[1]でそれぞれ 9 月 18 日と 21 日に公開された，エンターテインメントに関するオンラインニュース記事（Set1 と Set2）と，Lycos Japan[2]で 9 月 28 日に公開されたエンターテインメントに関するオンラインニュース記事（Set3）である．

提案手法のクラスタ生成手順は以下の通りである．

1. 各文書から，形態素解析ツール「茶筅 Ver. 2.02」[3]を利用して，固有名詞中心にキーワードを抽出する．

2. 抽出したキーワードを元に，前述のキーワード活性度計算アルゴリズムに基づいて活性伝搬を行い，1500 ステップ後に高活性化しているキーワード集合を抽出する．

3. 抽出した高活性化キーワードそれぞれについて，それを含む文書でクラスタを生成する．

ステップ (2) において，各キーワードの活性値は周期的に変動するが [11]，約 1000 ステップ経過後は定常状態と見なすことができた．

k-means の実行には，STATISTICA2000 (StatSoft, Inc.) を用いた．k-means では実行前にクラスタ数を指定する必要があるが，提案手法で生成されたクラスタと同数を指定して実行した．k-means の場合，各クラスタサイズはばらつきが大きく，一つの文書しか含まないクラスタも生成されることがあるが，それについては除外している．

大学の研究者および学生 9 人に，１人１データセットずつ，k-means によるクラスタリング結果と，提案手法による結果それぞれについて，各文書クラスタの結束性（話題に関するまとまり）を評価してもらった．具体的には，各クラスタ内の文書について,「互いに深く関連している」（評価点 5）から「関連していない」（評価点 1）まで，5 段階で評価してもらった．実験結果について，各データセット毎のクラスタ数，クラスタサイズの分散（サイズ分散），クラスタに対する評価の平均値（評価平均），評価（平均値）が 3.5 以上のクラスタ数（評価 A ），2.5 以上 3.5 未満のクラスタ数（評価 B ），2.5 未満のクラスタ数（評価表 C ）について，それぞれ表 2 に示す．これより，Set1, Set2 では提案手法の方が良い評価を得ていることがわかる．データセット毎の文書数，抽出キーワード数は，Set1 が 25 文書，75 キーワード，Set2 が 24 文書，62 キーワードであるのに対し，Set3 は 23 文書に対して抽出キーワード数が 22 と非常に少なく，これが Set3 のみ結果が芳しくない事に影響したものと考えられるが，今後検証，考察を進める必要がある．

k-means と提案手法で，同じ文書クラスタが生成されることも比較的多かった．また，クラスタサイズの分散が大きい事からもわかるように，k-means は一つの大きなクラスタを作る傾向があるため，提案手法で生成したクラスタがそれらの部分集合となる場合も多かった．そこで，k-means と同じクラスタ（KMEANS），部分集合，その他に分けて，評価点の分布を見た結果を表 3 に示す．これより，k-means と異なるクラスタであっても，十分よい評価が得られていることがわかる．

[1] http://news.yahoo.co.jp/
[2] http://www.lycos.co.jp
[3] http://chasen.aist-nara.ac.jp/index.html.ja

表 2: K-means との比較

データ	評価項目	提案手法	K-means
Set1	クラスタ数	5	4
	サイズ分散	**0.48**	3.6
	評価平均	**4.33**	3.90
	評価 A	**5**	2
	評価 B	0	1
	評価 C	0	1
Set2	クラスタ数	5	4
	サイズ分散	**0.32**	4.625
	評価平均	**3.82**	3.13
	評価 A	**4**	1
	評価 B	1	2
	評価 C	0	1
Set3	クラスタ数	5	5
	サイズ分散	**0.48**	4.25
	評価平均	2.3	**4.00**
	評価 A	1	**4**
	評価 B	1	0
	評価 C	3	1

WWW 情報可視化システム

情報可視化システムは，増大・普及の一途をたどる WWW 情報リソースの活用支援の一つとして有望視されている．ブラウジングを支援するための情報可視化システム [8] では，3D 表示の利用やインタラクティブ性の向上などにより，ページ間のリンク関係を把握しやすくするアプローチが多いのに対し，サーチエンジンを利用した情報検索支援には，クラスタリングを用いて検索結果文書集合の閲覧効率を高めたり [3, 16, 17]，キーワードマップを提示して検索結果中の話題分布を可視化し，次回以降の検索におけるクエリー生成を支援するシステム [9] などがある．

表 3: k-means 生成クラスタとの対応と評価点の分布

クラスタ区分	1	2	3	4	5	計
KMEANS	0	2	0	7	6	15
部分集合	2	1	0	6	0	9
その他	3	2	0	11	4	20
計	5	5	0	24	10	44

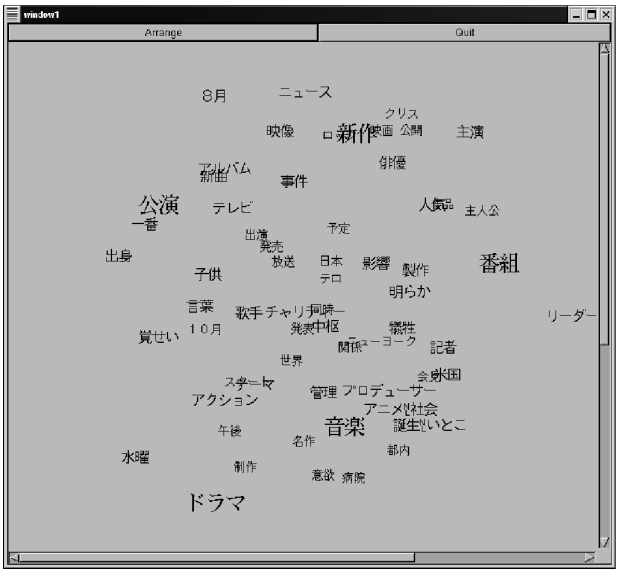


図 2: Set2 に関するキーワード マップ

また，文書クラスタリングが検索結果の閲覧を容易にする事が主目的であるのに対し，キーワード 間の類似性（距離）を元に，キーワードを二次元空間に配置する**キーワード マップ**は，文書集合に含まれる話題分布を提示するのに有効であり，テキストマイニング [15] や発想支援システム [14] などでも利用されている．本稿で提案する手法は，話題分布を反映してキーワード の活性度を計算するため，文書クラスタリングだけでなくキーワード マップ生成にも有効である．すなわち，ある話題に対応するキーワード 群は，クラスタ識別子が高活性化し，関連キーワードを抑制する形になっている [12]．従って，キーワード 活性度をキーワード サイズで表し，キーワード 間の距離だけを反映した通常のキーワード マップと組み合わせることにより，話題分布の状況をより明確にする事ができる．図 2 に，前述の実験で用いた Set2 の文書集合に対して作成したキーワード マップを示す．キーワード の空間配置には，バネモデル [14] を用いている．距離情報のみを反映したキーワード マップに比べ，キーワード 識別子（大きなフォントで表示）の周辺を関連キーワード（抑制され，小さなフォントで表示）が囲んでおり，話題分布の把握が容易であることがわかる．

図 3 に，現在開発中の，情報可視化プロトタイプシステムについて示す．提案システムは提案クラスタリング手法の特性を生かし，既存サーチエンジンからの検索結果を分析してキーワード マップを作成すると同時に，検索結果 (文書) をクラスタリングしてユーザに提示する事ができる．

プロトタイプシステムの特徴として，

- キーワード マップの可読性の向上
- 一連の検索結果間の関連性の考慮

があげられる．キーワード マップはキーワード 間の関係を 2(あるいは 3) 次元空間にお

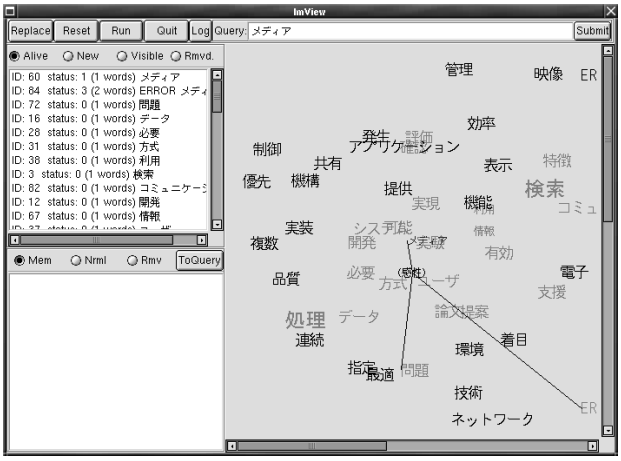


図 3: 情報可視化プロトタイプシステム

ける距離として表現するが，多数のキーワードが密集して表示される場合が多く，可読性が低下しやすい事，および解釈がユーザの主観に依存する度合いが高いという問題がある．プロトタイプシステムでは可読性を向上させるために，解釈の際のランドマークとすべき少数のキーワードの指摘，キーワード間の距離以外の特徴量としてキーワード活性度の提示 (図 3 ではキーワードのサイズで表現) を，図 2 と同様に行う．

本稿では詳しく触れなかったが，プロトタイプシステムでは異なる検索結果間の時系列的対応付けも考慮している．図 3 で薄い色で描画されたキーワードは以前の検索結果においても出現している事を示しており，またリンクを持つキーワード (図では「メディア」，「問題」など) は，リンク中央に示されたキーワード (図では「感性」) をクエリーとして検索した際にランドマークとして共起した事を示す．クエリーはユーザの知りたい話題 (親概念) に対応しており，その検索結果について共起するランドマークは，その話題を分割する話題 (子概念) と見なすことができる．従って，これらの情報を検索作業中のユーザに提示することにより，一連の検索作業を通じた，ユーザによる対象領域知識獲得を支援する効果が期待できる．

提案システムは，情報収集システムへも適用可能と考えている．検索過程にユーザが介入せず，自動的に行う情報収集活動では，現在までの収集状況を知ることが重要であり [6]，この目的のためにも文書クラスタリングを考慮して話題分布構造を抽出する事のできる提案手法は有効であると考える．

考察

WWW 上で多数公開されているオンラインニュース記事や，ユーザによる一連の情報検索作業により得られる検索結果など，WWW 上の情報収集においては時系列的関連を持った文書集合族を対象とすることが多い点に着目し，これを扱うための可塑的クラスタリング手法について提案し，生成されるクラスタの特性について，k-means との比較実験をアンケートに基づき行い，有効性を確認した．

提案手法は，文書集合中から抽出したキーワードを共起関係などに基づき接続したネッ

トワーク上で，免疫ネットワークモデルに基づく活性伝搬を行い，高活性化キーワードを抽出する．抽出したキーワードは，文書クラスタリングを行う際にクラスタ識別子として利用されるだけでなく，キーワードマップの可読性を高めるためにも利用可能である．また，文書集合間の時系列的関連性を考慮するために，免疫記憶細胞とのアナロジーを用いる．

提案手法の，WWW 上での情報収集／情報可視化システムへの応用の検討として，現在開発中の情報可視化システムのプロトタイプについても紹介した．今後は，被験者によるプロトタイプシステムの評価実験や，時系列文書集合族を対象とした評価実験を中心に行っていく予定である．

参考文献

[1] R.W. Anderson, A. U. Neumann, A. S. Perelson, "A Cayley Tree Immune Network Model with Antibody Dynamics," Bulletin of Mathematical Biology, Vol. 55, No. 6, pp. 1091–1131, 1993.

[2] C. Cole, "Interaction with an Enabling Information Retrieval System: Modeling the User's Decoding and Encoding Operations," Journal of the American Society for Information Science, Vol. 51, No. 5, pp. 417–426, 2000.

[3] M. A. Hearst and J. O. Pedersen, "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," SIGIR'96, pp. 76–84, 1996.

[4] N. K. Jerne, "The Immune System," Sci. Am., Vol. 229, pp. 52–60, 1973.

[5] 宮本，クラスター分析入門，第 2 章，森北出版，1999.

[6] 椰野，山田，マルチ Web ロボットによるユーザの興味を反映した情報収集，電子情報通信学会論文誌 D-I, Vol. J83-D-I, No. 7, pp. 780-788, 2000.

[7] A. U. Neumann and G. Weisbuch, "Dynamics and Topology of Idiotypic Networks," Bulletin of Mathematical Biology, Vol. 54, No. 5, pp. 699–726, 1992.

[8] 塩澤，西山，松下，協調検索型ハイパーメディアの WWW による実現，情処グループウェア研究会報告 95-GW-13, pp. 13–18, 1995.

[9] 砂山，大澤，谷内田，ユーザの興味の構造を用いて関連検索キーを提示する検索支援インターフェイス，人工知能学会誌，Vol. 15, No. 6, 2000.

[10] B. Sulzer et al., "Memory in Idiotypic Networks Due to Competition Between Proliferation and Differentiation," Bulletin of Mathematical Biology, Vol. 55, No. 6, pp. 1133–1182, 1993.

[11] 高間，廣田，クエリーのネットワーク化による話題分布構造可視化システムの構築，人工知能基礎論研究会資料 SIG-FAI-A003, pp. 13–18, 2000.

[12] Y. Takama and K. Hirota, "Employing Immune Network Model for Clustering with Plastic Structure," CIRA2001, pp. 178–183, 2001.

[13] Y. Takama and K. Hirota, "Consideration of Memory Cell for Immune Network-based Plastic Clustering method," InTech'2001, unpublished.

[14] 高杉，國藤，スプリングモデルを用いたアイデア触発のための思考支援システムの開発，人工知能学会誌, Vol. 14, No. 3, pp. 495–503, 1999.

[15] ビジュアルテキストマイニング, 人工知能学会誌，Vol. 16, No. 2, pp. 226–232, 2001.

[16] O. Zamir and O. Etzioni, "Grouper: A Dynamic Clustering Interface to Web Search Results," Proc. 8th Int'l WWW Conference, 1999.

[17] O. Zamir and O. Etzioni, "Web Document Clustering: A Feasibility Demonstration," Proc. SIGIR'98, pp. 46–54, 1998.

WWW 情報管理のための Web ページにおける部分情報の更新モニタリング

研究協力者　中井 有紀　(東京工業大学大学院総合理工学研究科)
研究代表者　山田 誠二　(東京工業大学大学院総合理工学研究科)

はじめに

現在インターネット，WWW の普及にともない，WWW を用いた情報収集が日常的に行われるようになった．WWW という情報源の大きな特長として，そこでの情報が時々刻々更新されることがあげられる．それにより，ユーザは毎日お気に入りの Web サイトを訪れ，最新情報の掲載された Web ページをチェックすることで，常に最新の情報を入手することができる．このように，Web の情報の多くは，それを一度収集すれば済むものではなく，収集した後に，更新をチェックして最新の情報を獲得するという **WWW 情報管理**が重要であり，一般にその管理はコストのかかるタスクである．この WWW 情報管理は，不特定・非定常・大規模・分散知識源の中から，ユーザの目的や興味に合致するデータやそれらの関連を効率良く探索し前処理する**アクティブ情報収集**を WWW において実現する場合に，特に重要となる．なぜなら，変動しやすいユーザの目的や興味に対して常に適合した情報を維持するには，更新された情報がユーザの目的に適合するものであるかどうかを常に監視する必要があるからである．

WWW には最新情報を提供するサイトが数多く存在するが，中でもユーザが頻繁に更新をチェックする必要のある Web サイトの例として，株価情報，天気予報，価格情報，ランキング，掲示板などのサイトがあげられる．ユーザは，これらのページを定期的にモニタすることにより，常に最新の情報を獲得できるが，反面これらの Web ページの更新は厳密には定期的に起こらない場合も多く，ユーザは最新情報を得るために，更新されていない Web ページを頻繁に見にいかなければならない．また，Web ページのチェックを怠ったために，大切な情報を取り逃がしてしまうこともある．

この問題を解決するために，Web ページ更新チェックアプリケーションが数多く存在している [15][2][13][11]．これらのアプリケーションは，ユーザの指定した Web ページで更新が行われた場合に通知するものであり，ユーザは更新されたページだけを見ることができ，更新情報を見逃す心配がない．しかし，問題は，ユーザが欲しい更新情報はページ全体ではなく特定の制約を満たす一部である場合が多いことである．株価であれば自分の持っている株の情報が分かればよく，天気予報であれば自分の地域の天気が分かればよいなどがそのような場合である．

図 1 に典型的な天気予報の Web ページを示す．例えば，ユーザが次の日曜日にハイキングに行く予定がある場合には，次の日曜日の天候のセルの内容が更新された場合だけに更新して欲しい．しかし，これまでの更新チェックツールでは，指定した Web ページにおいて，任意の更新が行われる度に逐一通知してしまう．Web ページ中のある制約を満たす部分だけの更新があった場合にだけ更新を通知することができない．我々は，このような更新を**部分更新**と呼ぶ．さらに，この部分更新は，どの部分にユーザが注目している

Weekly Weather
2001/1/1 17:00 JST

	Weather	Rain	High Temp			Low Temp		
			(F)	(C)	Diff	(F)	(C)	Diff
2002/01/03 (Thu)	cloudy, occasionally clear	30(%)	48	9.0	−2	32	0.0	−3
2002/01/04 (Fri)	cloudy	30(%)	51	11.0	0	33	1.0	−2
2002/01/05 (Sat)	clear, occasionally cloudy	20(%)	48	9.0	−2	37	3.0	0
2001/01/06 (Sun)	clear, occasionally cloudy	20(%)	50	10.0	−1	32	0.0	−3
2002/01/07 (Mon)	cloudy, passing rain	50(%)	46	8.0	−3	33	1.0	−2
2002/01/08 (Tue)	cloudy, occasionally clear	30(%)	48	10.0	0	34	0.0	−1

図 1: 天気予報の Web ページにおける表

かは，状況によって異るため，事前に特定することは難しく，その場その場でユーザと対話的に決定していく必要がある．

そこで，本研究では Web ページ上のユーザによって指定された一部分の情報に着目し，その部分に特定の更新があった場合のみ，その更新をユーザに提示する PUM(Partial Update Monitoring) システムを提案する [10]．ユーザに指定された部分の情報は，更新後の Web ページにおいても同じような位置，あるいは特定の条件に当てはまる位置にあると考えられる．一例としては，欲しい情報が表の中の同じ座標のセルにある場合や，表の特定の項目に注目して必要とする情報の場所を特定している場合があげられる．このような条件は，Web ページ中の HTML タグや表の見出し語，あるいはユーザに指定された部分の文字列等を用いたルールを使って表すことができる．しかし，このようなルールを記述することは，ユーザに相当程度の知識や労力を要求する．そのため，PUM ではユーザが監視させたい部分を特定するためのルールを帰納学習によって獲得する．また，特定された部分の更新がユーザにとって必要なものであるか判定するためのルールも学習する．

さらに，本研究では更新監視のためのルール学習をより効率的に行うため，ユーザから与えられる訓練例だけで学習を行うのではなく，システム側から例を提示しユーザに判定してもらう．また，更新結果だけでなくルールを直接ユーザに提示することも学習に有効である [12] と考える．

本研究と同じく Web ページの監視を行う研究として，新旧の Web ページの差分を検出し，その差分のみを提示する WebBeholder[13] があげられる．WebBeholder は，HTML タグにつけられた重みから求めた重要度が高い差分をユーザに提示する．また，Web Secretary

[1] はWeb上で公開されているWebページ差分通知システムであり，指定したWebページで更新された部分をハイライト表示する．しかし，更新結果として表示させる部分をユーザが決定しているのではない点でこれらのシステムと本研究とでは大きく異なる．

WWWは，膨大な大きさの極めて動的な情報源，あるいは知識源と考えられる．ただし，その知識源は，HTMLにより準構造化されてはいるが，ノイズを多く含んだ形式で記述されている．よって，その知識源からコンピュータ可読な十分に構造化された知識を抽出することが必要である．このような背景に基づいて，Webページからの様々な情報抽出に関する様々な研究が行われている [1][5][7][14] ．

HTML文書を生け垣として扱い，複数のHTML文書に対して反単一化を行うことで，一定の共通構造の中に含まれる情報を抽出する研究 [6] がある．また，松下らも，Webページの更新による差分に基づいて，情報抽出する方法を提案している [9]．これらの研究において，ユーザの必要とするデータが一定の規則を満たす場所に存在するという前提条件は本研究と共通しているが，本研究では対話的にユーザの意図するデータを求めている点で異なっている．

PUM

システムの概要

PUMの概要を図2に示す．PUMは，Webページ上でユーザの指定した部分を特定し，指定された部分にユーザが求める種類の更新があった場合に通知するWebページ更新監視システムである．

PUMは更新を監視するために，ユーザがWebページ上で指定した部分 (**指定部分**) から，**位置特定訓練例**，**更新判定訓練例**を取得し，帰納学習システムRIPPERを用いて**位置特定ルール**及び**更新判定ルール**を学習する．ルールが生成されると，Webページで更新があった場合，位置特定ルールで表中のセルなどの監視する領域 (**監視領域**) を特定し，更新判定ルールによって通知すべき更新かを判断する．

位置特定ルール，更新判定ルールにより更新が判定できなかった場合，訓練例を提示しユーザに評価させる，または再度指定部分を入力させることで，訓練例を追加し再び学習を行う．また，PUMはユーザとのインタラクションをもち，ユーザは任意のタイミングでルールそのものを評価することもできる．

更新通知手続き

PUMのWebページ更新通知手続きを説明する．なお，以下の手続きの番号と図2の番号が対応している．

1. 更新監視領域の取得.
   (a) ユーザは，Webページ中でPUMに監視領域を指定.
   (b) PUMは，監視領域を取得，解析する.
2. 更新監視ルールの作成.

[1] http://homemade.hypermart.net/websec/

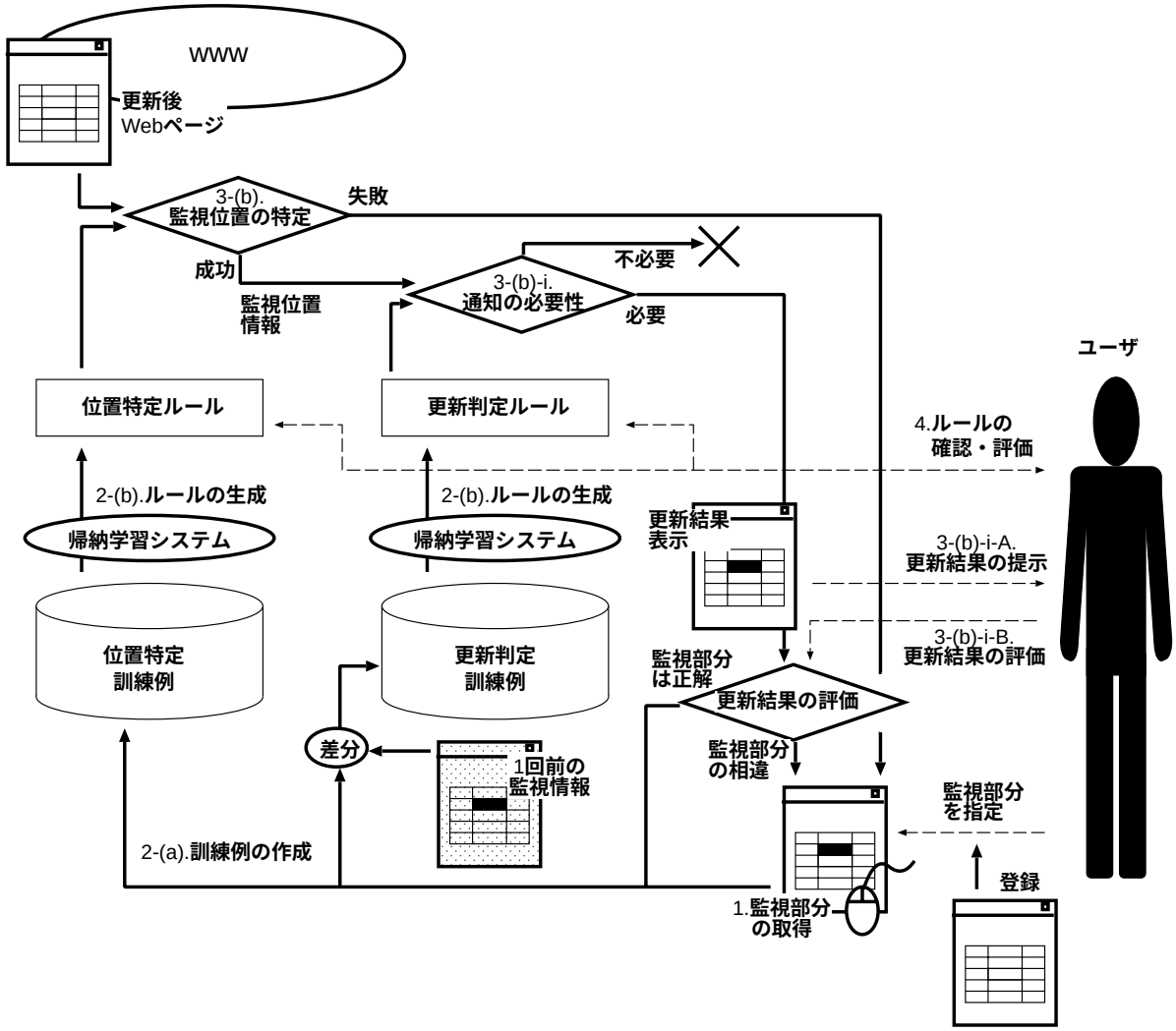


図 2: システムの概要

(a) 訓練例の生成

位置特定訓練例 以下に示す情報を持つ.

- 指定部分の HTML コード.
- HTML 構文ツリー上で指定部分の親にあたるタグの配列.
- 行，及び列の見出し語 (指定部分が表の中に存在する場合のみ).

更新判定訓練例 前回と今回の更新から得られた指定部分の情報の差分から作成する.

(b) それぞれの訓練例を独立に帰納学習システムに与え**位置特定ルール**，**更新判定ルール**を生成.

3. 更新監視ルールを用いた Web ページの監視.

(a) Web ページの更新を監視，更新有り→ (b)．更新無し→ (a).

(b) 位置特定ルールにより，監視位置を特定．特定成功→ i．特定失敗→ ii.

i. 更新判定ルールにより通知の必要性を判定．必要→ A．不必要→ C.

A. 更新部分をハイライトしたページをユーザに提示.

B. ユーザの評価.

- 監視位置は正しい→ 2.
- 監視位置が間違い→ 1.

C. 更新部分を提示しない．→ 3.

ii. 位置特定失敗．複数部分にマッチ→ A．一箇所もマッチせず→ 1.

A. すべての更新監視領域をユーザに提示.

B. ユーザの評価.

- 監視位置は正しい→ 2.
- 監視位置が間違い→ 1.

4. ユーザによるルールの評価（任意のタイミング）.

訓練例

PUM では更新監視ルールとして，監視領域を特定した後，ユーザに通知すべき更新か判定する．そのため，ルールも位置特定ルールと更新判定ルールの 2 つのルールを使用しており，それぞれのルールを生成するため，位置特定訓練例と更新判定訓練例の 2 つの訓練例を用意する.

本研究では，帰納学習システムとして，関係学習システム RIPPER [4][3] を用いる．RIPPER を採用した理由はいくつかある．まず，RIPPER はクラスの判別ルールを学習するため，本システムにおける重要な機能の一つであるユーザからの修正というインタラクションにおいて，学習された知識（ルール）の可読性が高いこと，そして，高い精度で効率のよい学習が可能なことがあげられる．RIPPER は，訓練例の属性値として，名義値 (nominal value)，集合値 (set value)，連続値を取ることができ，変数を含まない判別ルールが学習される．なお，RIPPER のソースコードは，Web ページ[2] から入手できる.

位置特定訓練例

Web ページ上でユーザが指定した部分，及びその部分に対するユーザの評価から，以下の形式で表される位置特定訓練例を作成する.

$$(tag1, tag2, tag3, ..., cNo, rNo, data, cIndex, rIndex, class)$$

まず，Web ページ上でユーザが指定した部分から，指定範囲の HTML コードをタグ，タグの属性，属性値，文字列，文字列を区切り文字で分解したもの，を $data$ とする.

[2]http://www.research.att.com/ diane/ripperd.html

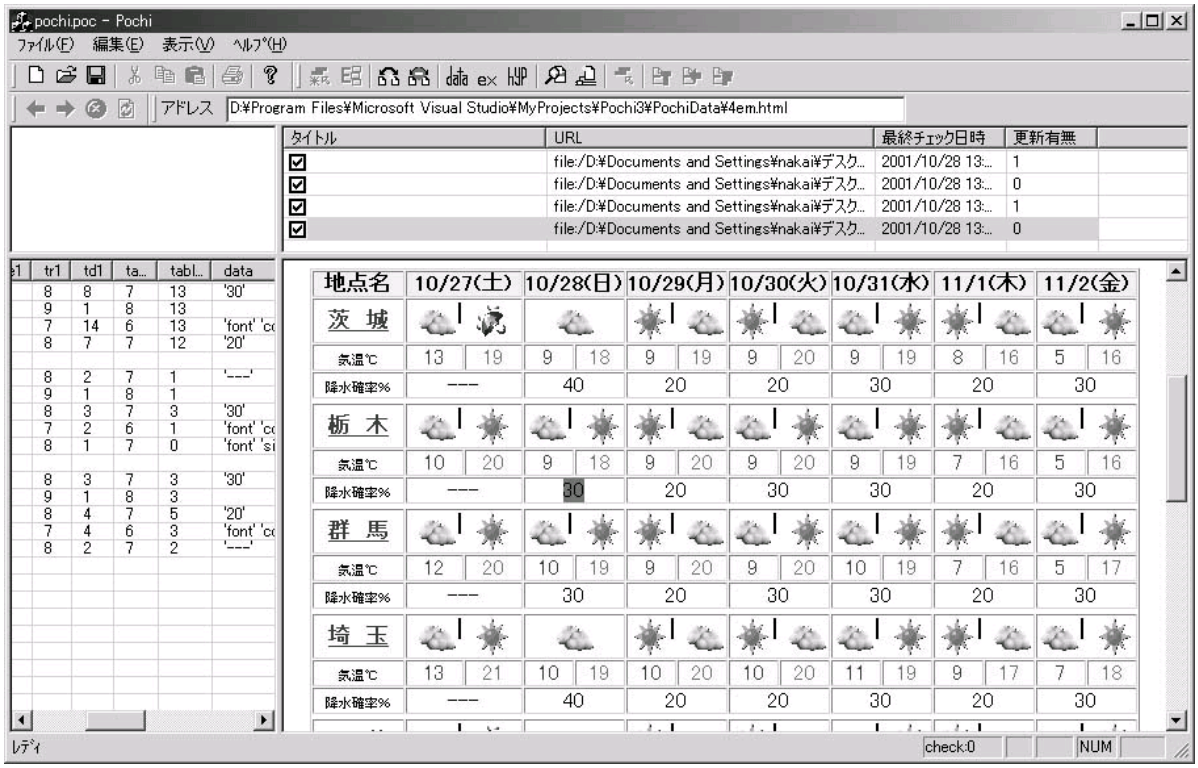


図 3: PUM のインタフェス

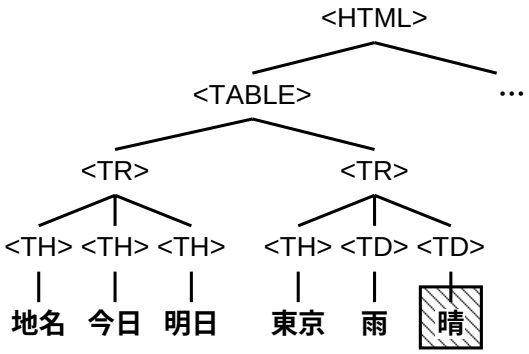


図 4: HTML 構文ツリー

さらに，指定された Web ページを HTML 構文解析することにより，HTML 構文ツリーを生成し，指定範囲を HTML 構文ツリー上にマッピングする (図 4)．そして，HTML 構文ツリー上で指定範囲の祖先に当たるタグのリスト $(tag1, tag2, tag3, ...)$ を求める．このとき，ここのタグについて HTML 構文ツリー上で兄に当たるタグの数を属性値とする．

指定範囲の祖先に当たるタグの配列は，更新されたページと更新前のページでは異なる可能性がある．そのため，PUM では LCS(Longest Common Subsequence)[8] を用い，異なる 2 つのタグのリストを包含する最短の 1 つのタグのリストを求め，そのタグのリスト

html	body	center	table	tr	td	rNo	cNo	data	cIndex	rIndex	class
1	1	1	2	8	8	7	13	'30'	'10/14(日)' '10' '14' ' 日'	'font' 'size' ' "−2"' ' 降水確率%'	Good.
1	1	1	2	8	7	7	11	'20'	'10/13(土)' '10' '13' ' 土'	'font' 'size' ' "−2"' ' 降水確率%'	Nogood.
1	1	1	2	8	9	7	15	'30'	'10/15(月)' '10' '15' ' 月'	'font' 'size' ' "−2"' ' 降水確率%'	Nogood.

表 1: 位置特定訓練例

を祖先のタグのリストとして用いる.

さらに，指定範囲が表の中である場合には，次の優先順位に従って列（行）の見出し語である $cIndex, rIndex$ を求め，$data$ と同じ手順で分解する.

1. 同じ列（行）にある<TH>タグで作られたセルの要素.

2. 同じ列（行）の 1 行（列）目のセルの要素.

また，左上のセルを $(0, 0)$ として，指定範囲がある表の行番号，列番号である cNo, rNo をもとめる.

位置特定訓練例の $class$ 属性は，その情報に対するユーザの評価である．$class$ は以下の 2 つの値をとる.

good 監視場所が正しい情報

nogood 監視場所が間違っている情報

図 4 で網かけされた部分が指定範囲である場合，位置特定訓練例の属性とその値は以下のようになる.

$$(html, table, tr, td, cNo, rNo, data, cIndex, rIndex, class)$$
$$= (1, 1, 2, 2, 2, 1, \text{晴}, \text{明日}, \text{東京}, good)$$

また，同じ表の中で指定範囲の近くに現れるデータは，指定範囲のデータと類似したものでありながらユーザには選択されなかったと考える．そのため，ユーザに選択されなかった指定範囲の上下左右 4 つの情報を負例として用い，境界条件の学習を促す.

位置特定訓練例は，ユーザが監視領域の指定，あるいは，更新通知結果の評価を行うごとに新たな訓練例が追加される．また，位置を特定する際には，位置特定ルールと位置特定訓練例の中のすべての正例に共通する属性値を用いる.

更新判定訓練例

前回の更新で得られた情報と今回の更新で得られた情報から，以下の形式で表される更新判定訓練例を作成する.

$$(dataN, dataN\_O)$$

教示回数	位置特定ルール		
1回目	Good	cIndex ~'10/14(日)', rNo ~'7'.	(2/0)
	Nogood	.	(8/0)
2回目	Good	rNo ~'7', cIndex ~' 日'.	(3/0)
	Nogood	.	(12/0)

表 2: 位置特定ルール

更新をユーザに通知する必要の有無は，ユーザの指定した範囲のデータに依存すると考えられる．そのため，PUM では更新情報のうち位置特定訓練例で定義した *data* の値と，その *data* の更新前後の差分を利用する．

更新判定訓練例の *class* 属性は，その情報に対するユーザの評価である．*class* は以下の 2 つの値をとる．

good 通知して欲しい更新

nogood 通知して欲しくない更新

PUM の実行例 1

実行例として，天気予報ページ（図 3）における部分情報の更新の監視を取り上げる．このページは週間予報のページであり，翌日から 1 週間分の天気が掲載されている．

ユーザの期待する更新通知

この例においてユーザが期待する更新は以下の通りである．

監視範囲 栃木の日曜日の降水確率

更新通知条件 降水確率が 30 %以下

ユーザは図 3 の網掛け部分を指定し，現在のデータが通知する必要のある情報であるかないかを指定した．

帰納学習

位置特定ルール

位置特定訓練例の一部を表 1 示す．表 2 にユーザが教示した回数（登録時：0 回目）とその教示後に得られた位置特定ルールを示す．1 回目の教示は，登録後初めてページが更新されたときに，PUM が持っていた位置特定ルールでは監視領域の位置が特定できなかったために行われた．このルールは,「10/14(日)」を含む見出し語を持つ，7 行目（日付は 1 行目）のデータを監視することを意味する．これ以降，4 日間は 1 回目の教示で得られたルールで位置を特定することができた．

しかし，5 日目にこのルールでは監視領域を特定することができなくなり，ユーザに教示を要求した．これは，週間予報表の日付が「10/14(日)」から「10/21(日)」に変わった

dataN	dataN_O	class
'30'	'<->'	Good
'20'	'<->'	Good
'30'	'<+>'	Good
'30'	'<=>'	Good
'20'	'<+>'	Good
'—'	'20' '—'	Nogood
'50'	'—' '50'	Nogood
'40'	'<->'	Nogood

表 3: 更新判定訓練例

教示回数	更新判定ルール		
1 回目	Good	.	(1/0)
2 回目	Good	.	(2/0)
⋮	⋮		
6 回目	Nogood	dataN ~'—'.	(1/0)
	Good	.	(5/0)
7 回目	Nogood	dataN ~'50'.	(1/0)
	Nogood	dataN ~'—'.	(1/0)
	Good	.	(5/0)

表 4: 更新判定ルール

ためである．そして，再学習により得られた位置特定ルールでは,「10/21(日)」ではなく「日」が使われている．このルールを用いれば再び日付が変化しても，監視領域の特定が可能になった．

更新判定ルール

表 3 に更新判定訓練例，表 4 にユーザが教示した回数（登録時：0 回目）とその教示後に得られた更新判定ルールを示す．<>の中は数値の比較結果を表している．

6 回目の教示で，初めて *Nogoo* が指定され，更新通知しない値がルールとして得られた．このように，更新判定ルールはとりうる値のすべての組み合わせをユーザが評価しないと確定しない．

PUM の実行例 2

PUM による部分更新検出は，基本的には，表形式に適用可能である．先の例で示した単一の表であれば，その内容に依存しない．例えば，図 5 のような表から構成される株価の Web ページにおいて，ある銘柄の株価をモニタリングして更新を検出，通知すること

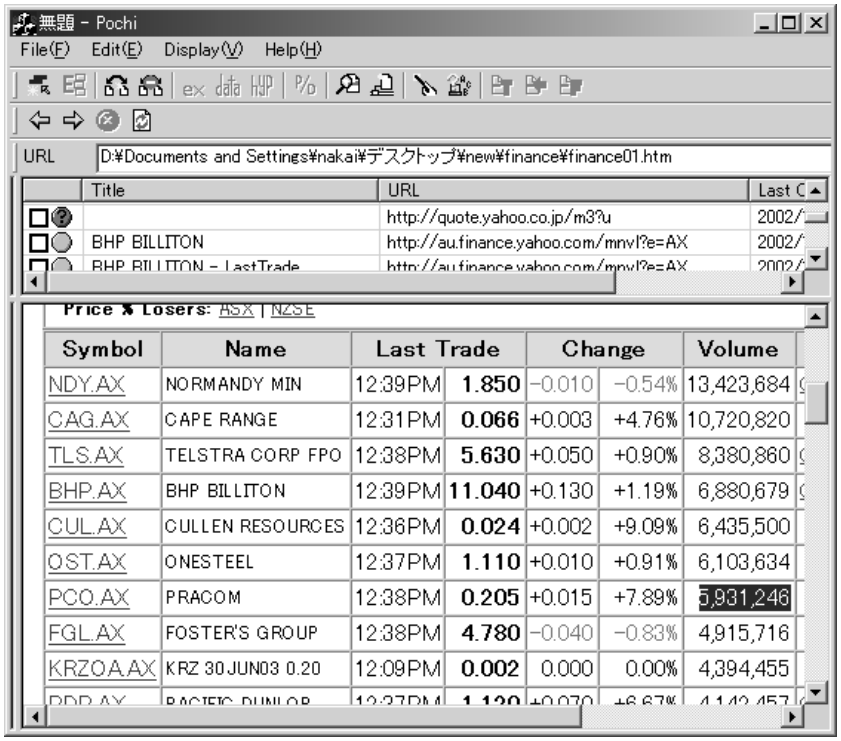


図 5: 株価の Web ページ

ができる．

さらに，図 6 のような 2 つの表にわたってセルがスクロールしながら更新が行われるような Web ページにおいても，同様に PUM による部分更新検出が適用可能である．

今後は，理論的，あるいは実験的に，PUM のリスト構造，プレーンテキストなどへの適用可能性を調べていきたい．

考察

ユーザとのインタラクション

ユーザにとって教示はコストがかかるため，ユーザの教示回数は必要最小限にとどめたい．そこで，ユーザとのインタラクションをもち，ユーザの知識を直接教示してもらうことにより，ユーザの教示回数，つまりユーザの負担を減らすことができる．PUM ではルールによる特定ができなかった場合に，ルールにマッチする候補をユーザに見せ選択してもらうことでより効果的な訓練例を獲得するインタラクションを実装している．

しかし，前章の実行例で示したように，PUM は位置特定ルールでは負例を自動生成することができるが，更新判定ルールでは負例を自動生成することが難しい．そのため，更新判定ルールでは学習が進みが遅い．更新判定ルールはシステム側で負例を生成するのが難しい反面，ユーザにとっては理解しやすく，ルールそのものをシステムに教示することも難しくないと考える．ユーザはユーザ自身が選択した範囲の情報に対する知識を持っており，その情報のとりうる値，値の範囲，更新パターンなどあらかじめ知っている．その

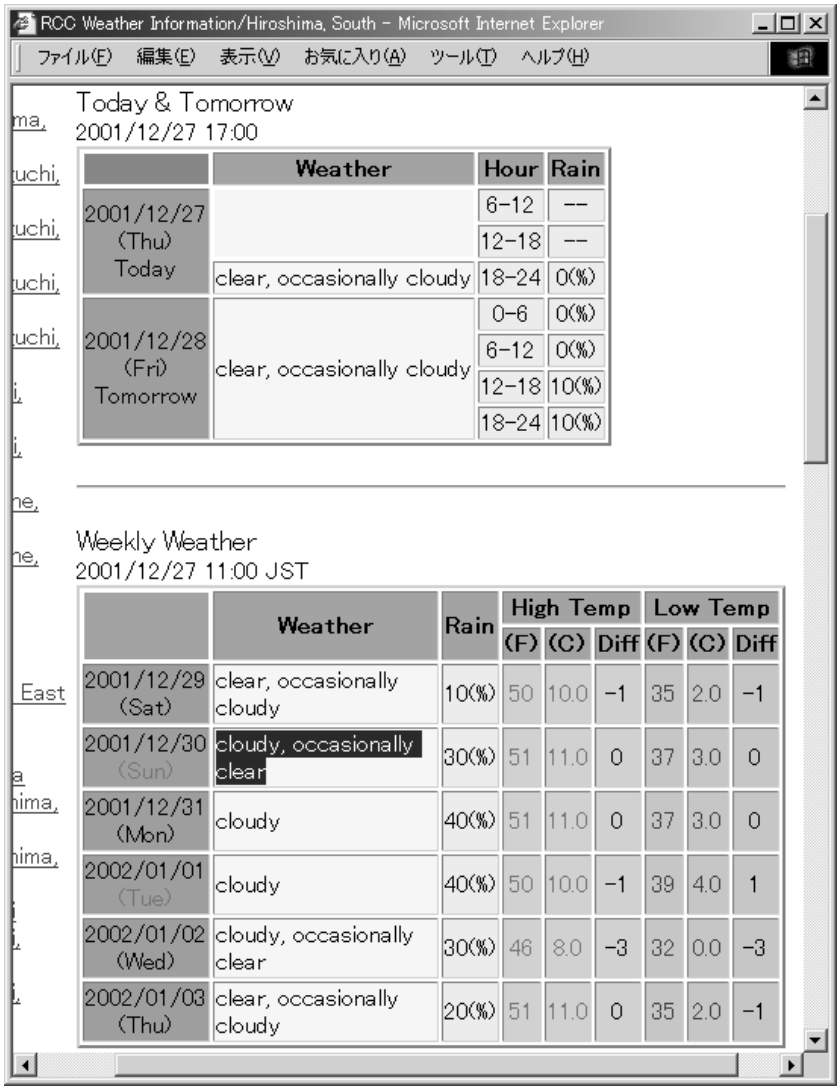


図 6: 2 つの表からなる天気予報ページ

ため，PUM 側がユーザにルールの教示を要求することが有効であると考えられる．

PUM の拡張

PUM では，複数の監視領域が特定の条件を満たした場合に提示することを考えてる．現時点では，位置特定ルールをそれぞれの監視領域に対して求め，それぞれの監視領域の位置を特定した後，前述の更新判定ルールを監視領域間の関係含む形に拡張した更新判定ルールを用い，ユーザに提示すべき更新であるか決定している．しかし，この方法ではそれぞれの監視領域が独立に決定する必要があり，監視領域間の関係が相対位置で表されるような場合には位置の特定が困難になる．また，このように複数の監視領域が相対位置で定義されるような場合は多く存在するため，この問題は解決する必要がある．

そのためには，位置特定ルールに各監視領域間の関係を記述すればよいが，監視領域間の関係をどのような形式で表記するかが，難しい問題である．

また，本稿では監視範囲が表である場合を扱ってきたが，監視範囲が表でないときの負例の取り方等システムの拡張が課題である．

まとめ

本稿では，Web ページ上でユーザが意図する部分，及び更新を関係学習によって学習し，ユーザの意図した更新が生じた場合にのみ更新情報を提示する，Web ページ更新監視システム PUM を提案した．ユーザは，自分の注目しているセルをマウスで領域指定すること，さらに PUM の通知してきた結果の評価をすることにより，PUM は，それらのユーザからの入力を訓練例として関係学習を行い，注目すべき領域の同定とその内容の満たすべき制約を学習し，自動的にユーザの意図する部分更新を検出，通知可能である．

今後は，PUM の適用範囲を明確にすることと，複数の監視領域や表以外の監視領域への拡張．また，ユーザに対するインタラクションのタイミング，表示方法を検討していく予定である．

参考文献

[1] M. E. Califf and R. J. Mooney. Relational learning of pattern-match rules for information extraction. In *Working Notes of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pages 6–11 (1998).

[2] ChangeDetection.com. http://www.changedetection.com/monitor.html.

[3] W. W. Cohen. Fast effective rule induction. In *Proc. 12th International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann (1995).

[4] W. W. Cohen. Learning trees and rules with set-valued features. In *Proceedings of AAAI'96*, pages 709–716 (1996).

[5] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and Seán Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118((1-2)):69–113 (2000).

[6] 福田，石野，竹田，松尾. 生垣上の反単一化を用いた情報抽出手法の提案. 第 54 回人工知能学会知識ベースシステム研究会, pp.47–52 (2001).

[7] Nickolas Kushmerick, Daniel S. Weld, and Robert B. Doorenbos. Wrapper induction for information extraction. In *Intl. Joint Conference on Artificial Intelligence (IJCAI)*, pages 729–737 (1997).

[8] T. コルメン，C. ライザーソン，R. リベスト. アルゴリズムイントロダクション, volume 2. 近代科学社 (1995).

[9] 松下，笹野，前田. 差分による WEB ページからの情報抽出のための基礎検討. 第 54 回人工知能学会知識ベースシステム研究会, pp.103–108 (2001).

[10] 中井有紀，山田誠二：Webページにおける部分情報の更新モニタリング，第54回「知識ベースシステム」研究会，pp.73–78 (2001).

[11] OmniViewer. http://www.digiportal.com/.

[12] M. J. Pazzani. Representation of electric mail filtering profiles: A user study. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces*, pages 202–206 (2000).

[13] S. Saeyor and M. Ishizuka. WebBeholder: A revolution in tracking and viewing changes on the web by agent community. In *WebNet 1998* (1998).

[14] S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233–272 (1999).

[15] Web Secretary. http://homemade.hypermart.net/websec/.

分散動的情報源からのアクティブ情報収集

研究代表者　北村泰彦　（大阪市立大学大学院工学研究科）
研究分担者　平山勝敏　（神戸商船大学商船学部）

背景と目的

インターネット(Internet)は我々の生活を支えるインフラストラクチャの一つとして急速に社会に浸透しつつある．インターネットをベースとしたサービスの中でもWorld Wide Webシステムはその中でも最も人気が高く，学術研究，電子商取引，個人やグループなどによる情報発信など，さまざまな目的のために利用されている．Webシステムはいまや地球規模の知識ベースを構築しているといっても過言ではないであろう．

Webシステムの特徴は，従来の分散データベースシステムと異なり，情報源がボトムアップに構築されるところにある．情報発信者はコンピュータをインターネットに接続し，Webサーバを立ち上げるだけで，即座に世界に向けた情報発信が可能になる．このようにWebシステムは，集中的な管理機構無しに，膨大な数の情報源が自律分散的に連携しているシステムであるといえる．

一方で，一利用者の観点からは有用な情報がネットワークの中に埋没してしまい，容易に見つけ出すことができないという問題も引き起こしている．これに対処するための解決法として検索エンジンが開発されている．しかしながら検索エンジンはキーワード入力に対して，それを含むWebページのURLリストを出力するだけである．中には，膨大な数のWebページを出力したり，また多数の関係のないWebページを含んだりするようなものも存在する．今後は関連Webページのリストを返すだけでなく，そのWebページの内容を解析し，その中から有用な情報のみを利用者に提供するような（広い意味での）データマイニング機構が望まれる．しかしながら，このような機構を実現するためには以下のようなWeb情報源の特徴を考慮する必要がある．

- Web情報の記述形式は非定型である．現在，ほとんどのWebページはHTML (Hyper Text Mark-up Language)により記述されているが，HTMLではWebページをブラウ

ザで表示する際に必要な視覚的な構造を表現することが可能であっても，ページに記述されている情報の意味的な構造を表現することは困難である．そこで Web ページから HTML タグ構造を手がかりに必要な情報を抽出するラッパーが必要になる．しかし今後は，意味的構造も記述可能とする次世代の Web ページ記述言語である XML[1] や Web 情報の機械的処理を目的とした Semantic Web[2] の研究や導入が行われており，情報抽出ラッパーの開発はより容易になると考えられる．

- Web 情報源に蓄積されている情報は日々，急速な勢いで増加している．報道機関など多くの情報源が 1 日に数回の情報更新を行っている．また株価情報や道路情報などは数分毎に情報更新を行っているものも少なくない．データマイニング機構が情報収集し，何らかの情報を発見したとしても，その情報がすでに古いものであれば，それは利用者にとって有用であるとはいえない．したがって情報源の更新を考慮したデータマイニング機構の開発が重要になる．

われわれは以上のような Web 情報源の特徴を考慮しながら，膨大な数の動的な情報源の中から利用者にとって適切な情報源を選択し，その中から有用な情報を発見するアクティブマイニングシステムを開発しようとしている．アクティブマイニングシステムの構成は図 1 に示される[3]．

- アクティブ情報収集モジュールは動的で大規模なインターネット情報源からアクティブマイニングモジュールや利用者に対して必要な情報を監視・提供する役割を果たす．
- アクティブマイニングモジュールはアクティブ情報収集モジュールにより収集された情報を解析し，利用者にとって有用な情報を発見する．
- アクティブユーザリアクションモジュールは利用者とのインタフェースの役割を果たし，利用者の要求に変化が生じた場合はそれをアクティブ情報収集モジュールやアクティブマイニングモジュールに伝達する．

以上のような三つのモジュールが互いに連携しあうことによりアクティブマイニングが達成される．

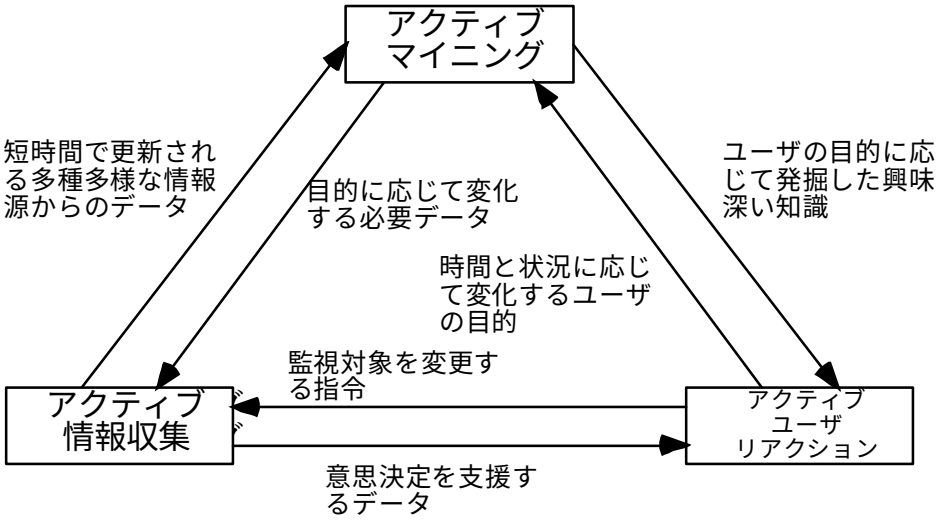


図 1：アクティブマイニングシステムの構成図

われわれはこの中でアクティブ情報収集に焦点を絞り，アクティブ情報収集システムに要求される機能と関連研究の調査，新たなアクティブ情報収集システムの構想とそのプロトタイプ作りを行う．

検討内容

アクティブ情報収集システムの機能と関連研究

アクティブ情報収集システムはインターネット上に存在する動的に変化するWeb情報源から利用者の要求を満たす情報を効率よく収集し，さらにその変化を監視するシステムである．アクティブ情報収集システムに求められる機能や関連研究は以下のようにまとめられる．

情報監視機構

Web 情報源の特徴の一つは情報源が頻繁に更新されることである．更新の頻度は情報源により異なるが，大学のホームページのように１週間に１度程度しか更新が行われないようなものから，新聞社のホームページのように１時間に数回の更新が行われるようなものもある．さらには，株価，スポーツ中継，オークション，道路情報などの情報提供サイトでは数分に一度の割合で更新が行われるものもある．利用者はこのよ

うに頻繁に更新される情報源から効率よく情報収集を行うことを望んでおり，そのような要求に応えるいくつかのシステムが開発されている．

AT&T で開発された AIDE (AT&T Internet Difference Enigne) [4]は WWW の変化を監視し，その違いを表示するシステムである．この中で二つの Web ページを比較し，その違いを表示する HtmlDiff (http://www.research.att.com/~douglis/aide/)と呼ばれるモジュールが開発され公開されている．またその改良を行い，Java により実装したものとして TopBlend[5]がある．

Lawrence Livermore National Laboratory で開発されている DataFoundry(http://www.llnl.gov/casc/datafoundry/index.htm)は情報源の変化を発見し，データウェアハウスのメンテナンスを行うシステムである．[6,7] ここでは科学データ源におけるデータベーススキーマをグラフ表現し，データとスキーマの変化の検知する．従来，科学データベースにおいてスキーマ変更は頻繁に行われるが，それを手作業で行うにはコストがかかっていた．情報源を定期的に監視し，自動的にスキーマを変更しようとする試みである．

また INRIA では XML ベースのデータウェアハウスのためのデータ監視システム Xyleme (http://www.xyleme.com/index.jsp)が開発されている．[8]

これ以外にも情報源監視を連続的なクエリ（continuous query）とみなすデータベースシステムからのアプローチとして Oregon Graduate Institute の CONQUER[9] や University of Wisconsin の NiagaraCQ[10] と呼ばれるシステムも開発されている．

差分表示機構

Web 情報源の変化が検知されたとき，利用者はその変化が知らされるだけでなく，どのように変化したかを分かりやすく知りたいという要求がある．このような目的にために前節で述べた HtmlDiff システムでは図２のように，二つの Web ページの違いを際立たせるために過去のデータには取り消し線を，新しいデータはイタリックで表示させるようにしている．

図 2：HtmlDiff の出力画面

評価機構

Web 上には同様の情報を扱う情報源が多数存在する．例えば，新聞社のホームページなどは多数あり，同様の情報を発信しているといえるが，その視点や頻度はそれぞれ異なっている．利用者はより早く，より有用な情報を入手したいと望むであろう．これは複数の情報源を監視し，更新の早さや量を比較することにより，情報源の近似的な評価を行うことが可能になると思われる．

統合機構

複数の Web 情報源から得られる情報を統合することはそれぞれの情報源の付加価値を高めることになる [11]． 例えば，新聞サイトから東海道新幹線が運休するニュースを入手した東京にいる旅行者が，航空会社のサイトから羽田空港発の航空便の空席情報を直ちに入手できるならば有意義である．また，同種の情報源からの情報を組み合わせることも有用である．例えば，同じ話題のニュースであっても，それが多くの情報サイトで取り上げられているとすれば，そのニュースがより重要であることが分かる．

このような情報統合機構を実現する場合には，情報収集の質とコストを考慮する必要がある[12]．例えば，情報の質を収集した情報源の数で評価するとするならば，より高い質の情報を得るためには，より多くの情報源から情報収集する必要があり，より多くのコストが必要になる．したがって一般には情報収集の質とコストはトレードオフの関係にあるといってよい．質とコストをうまくバランスをとりながら情報収集するためにはそのためのプランニング機構が必要になる．このような目的で Massachusetts 大学では BIG と呼ばれる情報収集エージェントが開発されている．[13]

アクティブ情報収集・統合システム Intelligent Ticker の設計

以上の調査を踏まえてわれわれは Intelligent Ticker と呼ばれるアクティブ情報収集・統合システムの設計を行った．

頻繁に更新される情報源からの情報収集を考えた場合，一時に更新されるそれぞれの情報量はそれほど大きくはない．例えば，新聞社のサイトにおいてそのトップページは数分の単位で頻繁に更新されるが，更新される量は Web ページ中の数行である．テレビ等で行われるニュース速報にしても数行のテキストが画面の上部に表示されるだけである．我々はこのように少量で速報性のある情報オブジェクトを Ticker と呼び，それらを Web 情報源から収集し，統合することで利用者の意思決定や問題解決を支援する Intelligent Ticker システムを提案する．

Intelligent Ticker は図 3 に示すように情報抽出部と情報統合部から構成される．情報抽出部は指定された Web 情報源に対して，その変化を監視し，その変化が存在した場合は，その差分のみを Ticker と呼ばれる情報オブジェクトとして生成する．

情報統合部は複数の情報抽出部で生成された Ticker を選択，組み合わせることにより利用者の意思決定や問題解決を支援する．利用者は情報統合部で得られた Ticker を直接表示させることも可能である．

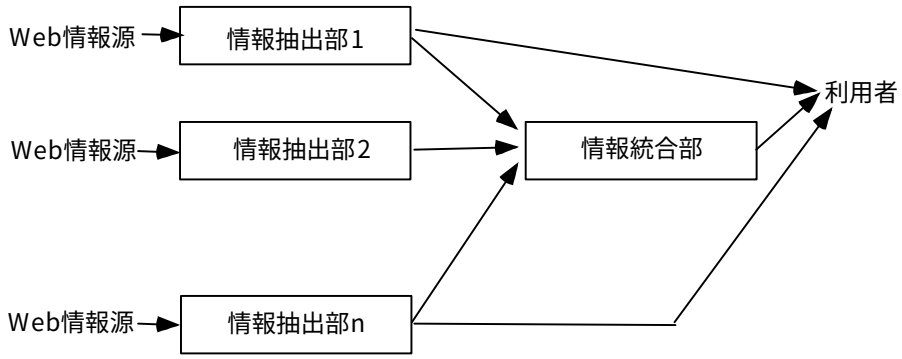


図３：Intelligent Ticker の構成図

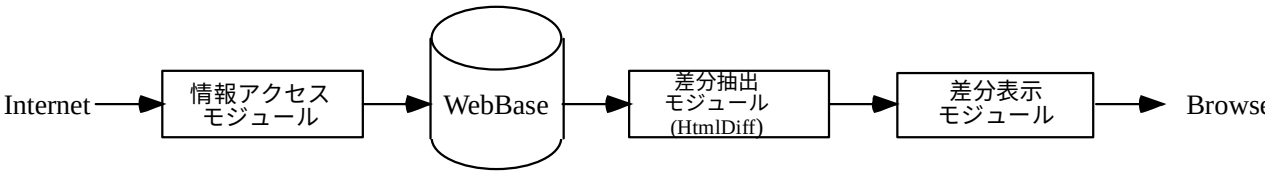


図 4:情報抽出部の構成

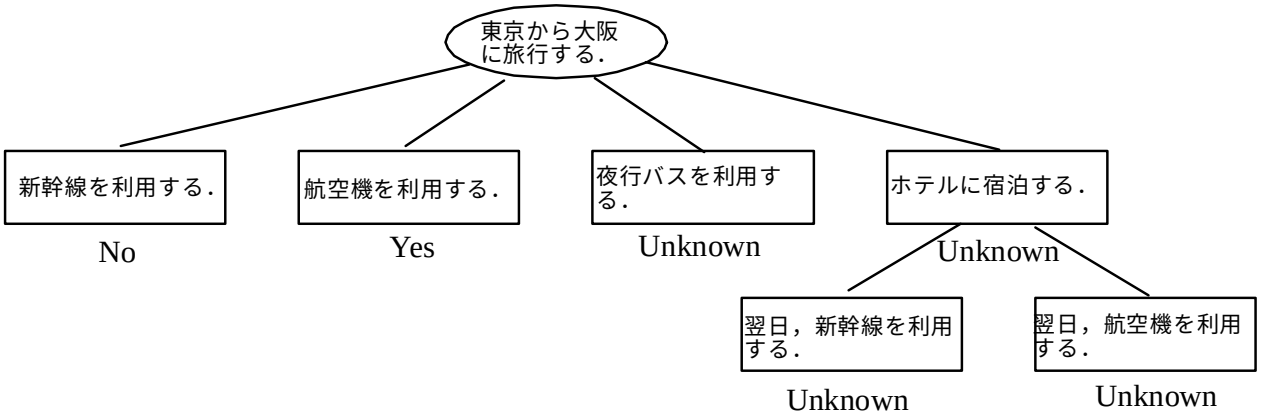


図 ５：Template for Ticker Integration

情報抽出部

情報抽出部の構成は図４のようになっている. 情報アクセスモジュールはインターネット上の情報源から定期的に Web ページをフェッチし，WebBase に保存する. 差分抽出モジュールは利用者から WebBase に格納されている二つの Web ページの指定に対してその違いを抽出する. このモジュールは先述の HtmlDiff を利用することができる. HtmlDiff はもともとの Web ページにタグを埋め込むことによりその違いを表示して

いるが，差分表示モジュールは変化した部分だけを利用者に対して表示する．
差分表示モジュールでは差分抽出モジュールにより抽出された差分を表示するが,差分のみを表示するだけでは，その文脈が取り去られてしまい差分の意味が理解しにくくなってしまうことが考えられる．そこで Web ページを構成する HTML 文書のタグの入れ子構造を解析し，差分の上位概念を示す構造は残すようにしている．
情報抽出部は情報源の変化に対して Ticker を生成する．Ticker は以下の要素から構成される．

- オブジェクト：更新された情報の断片そのもの．テキストやハイパーリンクなどにより表現される．
- タイムスタンプ：更新された時刻．
- ロケーション：更新された情報の URL．
- コンテキスト：更新された情報の文脈．

情報アクセスモジュールでは情報源の更新頻度を考慮した情報アクセスが望まれる．すなわち,例えば,1 日に 1 回しか更新されないことが分かっている情報源に対して，1 時間ごとに情報収集したとしても無駄である．したがって情報収集の操作を行いながら，情報源の更新頻度を学習し，それに応じて情報アクセスの間隔を変化させるような適応的な機能が必要になるであろう．

情報統合部
情報抽出部は情報統合部からの要求に対して，監視中の情報源に変化が生じたときに Ticker を発生するシステムである.情報統合部は複数の情報抽出部から送られてくる Ticker を統合して利用者の問題解決を支援する．
Ticker の統合には図５に示されるような TTI (Template for Ticker Integration)が用いられる．この図では具体例として，東京から大阪に旅行する場合のプラン候補が示されている．情報統合部は新幹線情報に関する Ticker と航空便情報に関する Ticker を選択しており，現在新幹線は満席で利用不可能，航空機は空席があり利用可能であるとする.ここで航空機が欠航になったことを通知する Ticker を受け取ると，新たに夜行バスに関する Ticker を収集し，それが利用可能でない場合は，ホテルと

翌日の交通手段に関する情報を収集する．もちろん，情報収集の途中で新幹線に空席が生じれば，夜行バスやホテルに関する情報収集は停止してもよい．このように情報統合部では得られる Ticker の内容に応じて動的に情報収集の方法を変更してゆく．
以上のような情報収集を行う際に情報を静的なものと動的なものに区分することは有効である．ここで動的な情報とは頻繁に更新される情報を指し，新幹線や航空便の空席情報に該当し，情報収集の主な対象となる．また，静的な情報とは動的な情報を収集するために利用される比較的変化の少ない安定した情報を指し，旅行プラン知識に該当する．静的な情報を用いて情報収集を行うことによって，やみくもに情報収集を行うのではなく，効率的な情報収集が可能になる[12]．
ここでは便宜上，情報を静的なものと動的なものに明確に区分したが，実世界では完全に静的な情報とはごくわずかであろう．例えば，上記の旅行プランに関しても，夜行バスの廃止や，東京大阪間の豪華船クルーズといった新たな交通手段の出現の可能性もある．すなわち静的な情報も長期的な視点からは更新を行ってゆく必要がある．そこで収集した情報をもとに動的な情報だけでなく，静的な情報も更新してゆく仕組みは必要であり，今後はそのための手段としてデータマイニングの技術は重要になると考えられる．

結果

情報抽出プロトタイプ：HtmlDiff を用いた差分抽出

Intelligent Ticker における情報抽出部のプロトタイプを開発した．プロトタイプでは朝日新聞の二つの Web ページ入力に対して図２に示すように HtmlDiff を用いてその差分を抽出し，その差分の中で特に意味をもつ構造だけを抽出して表示するようにしている．その出力結果を図６に示す．抽出すべき有効な構造を選択するためには図７に示すように Web ページを木構造に解析し，変化のあった部分とその上位構造を抽出するようにしている．ただし図６からもわかるように不必要と思われる広告バナーが表示されており，重要な部分をいかにして自動的に抽出するかは今後の課題となっている．なお，Ticker は図６で示される情報をさらに細分化したものとして生成される．

情報統合プロトタイプ：航空便空席照会システム

静的な情報と動的な情報を用いた情報統合部のプロトタイプとして航空便空席照会システムの開発を行った．本プロトタイプの特徴は以下のとおりである．

- インターネット上に存在する国内航空三社のホームページから航空便の空席情報を収集する．この実現にはクエリ（搭乗日，出発空港，到着空港）の送出と結果ページから空席に関する情報抽出を行うラッパーを Java により記述している．
- 空席照会は出発地から到着地への直行便だけでなく，収集した情報を統合することにより，乗り継ぎ便に関する情報も提供する．またこの乗り継ぎは異なった航空会社間の乗り継ぎも扱っている．
- 情報収集には静的な情報を利用している．ここでの静的な情報は航空機の乗り継ぎ経路である．乗り継ぎ便も含めるとシステムは何度も情報検索を行う必要があるが，利用者が入力する希望到着時刻にできるだけ近く，また飛行時間が短い便の優先順位を高くして情報収集を行うようにしている．

プロトタイプの出力結果を図 8 に示す．ここでは大阪（伊丹）発から札幌行の航空便を希望到着時刻 17 時として検索し，この図はプロトタイプが航空会社のホームページに 3 回の検索の後に得られた結果である．希望到着時刻に近く，旅行時間が短いものが上位に表示されている．

考察

ここでは今後の課題について考察する．

情報収集プランニング

これまで航空機の乗り継ぎ問題という具体例を対象に情報収集プランニング機構を開発してきたが，今後のその一般化を行い，他の領域の問題にも応用可能なものにする必要がある．問題が大規模になると問題解決に必要なすべての情報をあらかじめ収集してから解決を開始するのは非現実的である．現実的には部分的な情報を収集し，その中から解を導き出す必要がある．しかし収集する情報が少ないと，良い解が得ら

3回目 %Resolution: 180msec

75	JL 0577	ITM (1455) → (1640)SPK
71	JL 0106 ANA 069	ITM (1430) → (1535)HND HND (1600) → (1730)SPK
70	JC 0815 JC 0847	ITM (1410) → (1520)SDJ SDJ (1545) → (1700)SPK
61	JL 0104 ANA 067	ITM (1255) → (1400)HND HND (1500) → (1630)SPK
37	JL 0102 ANA 061	ITM (0855) → (1000)HND HND (1100) → (1230)SPK
37	JL 0573	ITM (0930) → (1115)SPK
30	JC 0811 JC 0843	ITM (0800) → (0910)SDJ SDJ (0935) → (1050)SPK
30	JL 0100 ANA 059	ITM (0720) → (0825)HND HND (1000) → (1135)SPK
29	JL 0100 JL 0507	ITM (0720) → (0825)HND HND (0950) → (1120)SPK
29	JL 0571	ITM (0820) → (1005)SPK
27	JL 0100 ANA 057	ITM (0720) → (0825)HND HND (0850) → (1025)SPK

図 8：航空便空席照会システムの出力結果

れない場合もある．したがって解の質と収集する情報の量の間にはトレードオフの関係があると考えられる．したがって今後は，利用者の課す時間的制約のもとで，効率よく情報を収集しながらより良い解を導き出す機構の開発が必要になる．このために制約充足問題解決の観点から問題の定式化を行う予定である．

静的な情報の更新

プロトタイプシステムでは航空機の乗り継ぎ経路は静的な情報として利用され，更新されることは前提としていなかった．しかし，航空機のダイヤや運行経路は毎月更新される可能性がある．したがって収集した情報をもとに静的な情報の更新を行うことも重要である．この機構を組み入れることにより情報収集のためのメンテナンスコストの軽減が期待される．また単なる更新だけでなく，収集した情報をもとに新たな乗り継ぎ経路を自動的に発見することも可能になるかもしれない．このように収集した情報から新しい静的な情報を発見するためにデータマイニングの技術を応用することができると期待される．

EBM への応用

アクティブ情報収集システムの他分野への応用として医学分野への応用を検討する．現在の EBM を支援するインターネット上の医学情報源は MEDLINE など主にテキストベースのものが多く，直接計算機で処理することは必ずしも容易ではない．そこである程度分野を限定する必要がある．そこで構造化された情報を提供する医学分野として医薬品の副作用情報の提供への応用を検討する．現在，医薬品情報に関しては厚生労働省を中心にその整備が行われようとしている[14]．医薬品はどんなに有効性の高いものであっても，その利用法に対する情報が備わっていなければ，その有効性が減少するだけでなく，逆にその危険性も生じかねない．近年では市販本の出版にも見られるように，医薬品に関する情報は医師や薬剤師などの医療専門家だけでなく，一般市民からの要求も増えている．医薬品情報は副作用など生命にかかわる重要なものであるとともに，次々と新薬が開発される現在では頻繁に更新される情報の一つである．また治療に対して複数の医薬品が用いられることは一般的であり，その組み合わせによる副作用も無視できない．このように動的な情報源から複雑な要求を満たす情報検索を支援するためにアクティブ情報収集システムの応用は有効であると期待される．

参考文献

[1] Klein, M.: XML, RDF, and Relatives. IEEE Intelligent Systems 16:2, 26-28 (2001)

[2] Fensel, D., Musen, M.A.: The Semantic Web: A Brain for Humankind. IEEE Intelligent Systems 16:2, 24-25 (2001)

[3] 元田浩：情報洪水時代におけるアクティブマイニングの実現，科学研究費補助金「特定領域研究(B)」申請書 (2001)

[4] Douglis, F., Ball, T., Chen, Y.-F., Koutsofios, E.: The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web. World Wide Web, 1: 27-44 (1998)

[5] Chen, Y.-F., Douglis, F., Huang, H., Vo, K.-P.: TopBlend: An Efficient Implementation of HtmlDiff in Java. In WebNet'00 (2000)

[6] Adam, N., Adiwijaya, I., Critchlow, T., Musick, R.: Detecting Data and Schema

Changes in Scientific Documents. In IEEE Advances in Digital Library (2000)
[7] Critchlow, T., Fidelis, K., Ganesh, M., Musick, R., Slezak, T.: DataFoundry: Information Management for Scientific Data. IEEE Trans Inf Technol Biomed, 4(1): 52-57 (2000)
[8] Nguyen, B., Abiteboul, S., Cobena, G., Preda, M.: Monitoring XML Data on the Web. In ACM SIGMOD (2001)
[9] Ling Liu, Calton Pu, Wei Tang, and Wei Han. Conquer: A continual query system for update monitoring in the www. International Journal of Computer Systems, Science and Engineering, 14(2): 99-112 (2000)
[10] Jianjun Chen, David DeWitt, Fend Tian, and Yuan Wang. NiagaraCQ: A scalable continous query system for the internet databases. ACM SIGMOD, 379 (2000).
[11] 山田誠二，村田剛志，北村泰彦．知的 Web 情報システム，人工知能学会誌，16(4):495-502 (2001)
[12] 北村泰彦，野田知哉，辰巳昭治．動的情報メディエータのための知的情報収集手法，電子情報通信学会論文誌 D-I，J84-D-I(8):1256-1265 (2001)
[13] Lesser, V., Horling, B., Klassner, F., Raja, A., Wagner, T., Zhang, S.X.: BIG: An agent for resource-bounded information gathering and decision making. Artificial Intelligence, 118(1-2): 197-244 (2000)
[14] 医薬品情報提供のあり方に関する懇談会最終報告～医薬品総合情報ネットワークの構築に向けて～，http://www.mhlw.go.jp/shingi/0109/s0927-2.html (2001)

伝言ゲーム型の情報収集とデータ前処理

研究代表者	沼尾 正行	(東京工業大学大学院情報理工学研究科)
研究協力者	森山 甲一	(東京工業大学大学院情報理工学研究科)
	Tran Tuan Nam	(東京工業大学大学院情報理工学研究科)
	Cholwich Nattee	(東京工業大学大学院情報理工学研究科)
	吉田 匡志	(東京工業大学大学院情報理工学研究科)
	伊藤 雄介	(東京工業大学大学院情報理工学研究科)
	束野 真人	(東京工業大学大学院情報理工学研究科)

1 はじめに

データマイニングを行うためには、人手によりデータを収集し、それらをマイニングツールに適用可能な形に整理せねばならない。これら、データ収集および前処理については、サンプリングの技法などが研究されているが [11]、まだまだ未開拓の分野であり、アクティブマイニングでは、図 1 に示すように、これらを含めた工程全体を研究対象とする。従来は最終工程であるマイニングやその直前の前処理が主に研究されてきたが、それらよりも前の段階も対象としていくわけである。

情報収集および前処理は、情報提供者、ドメイン専門家、マイニング専門家の共同作業になることがほとんどで、それらの間で大量のデータが交換され、更新が頻繁に行われることに特徴がある。このような共同作業を支援するため、通常、メールと Web ページが用いられるが、不便な点が多い。データマイニングの作業はルーチンワークではないので、従来のグループウェアの適用も困難である。

筆者らは、データマイニングに限らず、一般的な情報交換のツールとして、口コミ支援システム [13, 14, 15, 23] を提案し、システムを構築して実験を進めている。これは自然言語、画像、URL より構成されるメッセージを交換することを目的に設計されたものだが、データマイニング用のデータの交換、評価、前処理結果および手順の交換、マイニング結果の交換、評価などにも便利に使えると考えている。

本稿では、口コミ支援システムの概要を述べ、データマイニングに適用するための改善点について、考察してみたい。

2 情報収集

計算機ネットワーク上には様々な情報が氾濫し、ユーザーにとって有用な情報を獲得するのに大変な労力を必要とする。情報を発見するための最も身近なツールとして、Yahoo や Goo、Google[10] といった検索エンジンがあるが、最もカバー率の大きいものでも、せいぜい 4 割から 6 割程度であると言われており、広大な WWW 空間に氾濫した情報をすべて把握するのは、非常に困難である。

そこで、情報の内容を解析し、ユーザーが文章中のどの部分に対して興味を持っているか推定することで、同様の部分を持つ文書を有用な情報として提供する Content based filtering[20, 5] や、評価傾向の似ている他ユーザーの持つ情報を参考にして、有用な情報

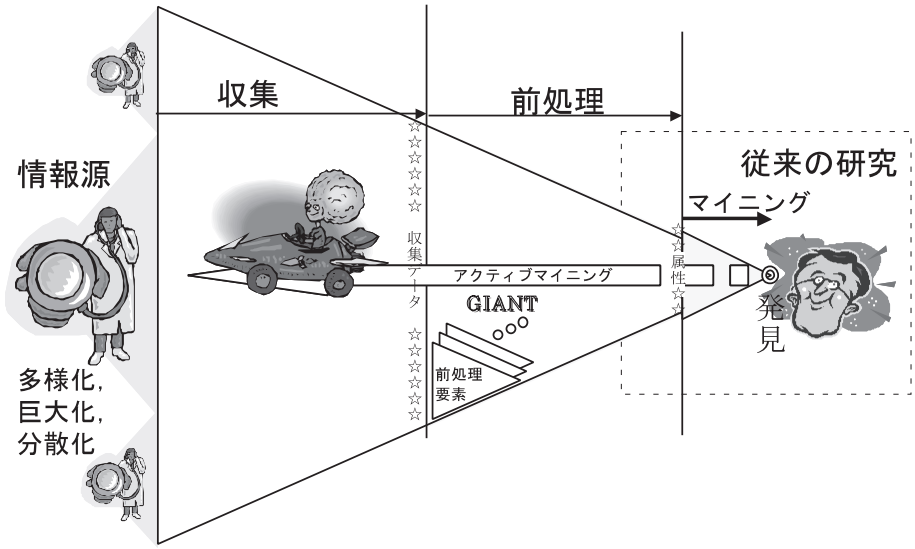


図 1: アクティブマイニング

を提供する Colaborative filtering[9, 22] などの情報フィルタリングの研究が盛んに行われている。

これらは結局、人をモデリングしていることになっており、かなり困難な問題を扱っている。人の内部を推定してモデル化するには、アンケートを取ったり、プロトコル解析するなどの観察に頼るか、脳波、MRI、光トポグラフィなどを用いることになる。いずれにせよ、リアルタイムでユーザをモデル化するには障害が多い。たとえ内部が完全に分かっても、社会レベルの活動への影響を解析するのは困難である。人にはいろいろな側面があり、ごく一部の側面のみしかモデル化できないからである (図 2)。

これに対して、大量の意見を要約して質の高い情報を提供するマスメディアと情報発信の自由やインタラクティブ性を持つ電子メディアを融合した新しいメディア [27] や、コミュニティの可視化 [28, 24, 21] など、人の心の動きや人間関係を考慮し、コミュニティの形成を支援する研究もある。しかし、情報フィルタリングのように、有用な情報獲得支援といったことまで考慮されていない。

本稿では、人間関係を電子コミュニティ上に再現することで、効率のよい情報収集や円滑なコミュニケーションを支援するシステムを提案する。その上に推薦機構を導入することにより、図 2 のように、人を独立した個体としてモデル化するのではなく、図 3 のように、現実に情報のやりとりが行われている人間関係をモデル化する。人の間のコミュニケーションには、ある程度の持続性があり、一貫したモードが存在する。それを解析した方が、個体をモデル化するよりも容易だと考えるからである。

3 口コミとは

口コミとは、「口から口へ伝えられる評判」(三省堂「大辞林」) のことであり、マスコミをもじって作られた言葉である。多くの人に均一に情報を伝えるマスコミに対して、口コミは、個人間で情報の伝達が行われ ([26] p.91)、図 4 のように、氾濫する情報の中から、

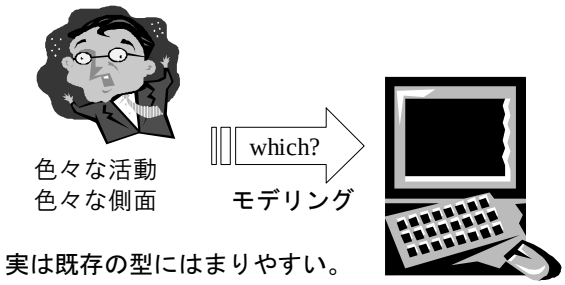


図 2: 人のモデリング

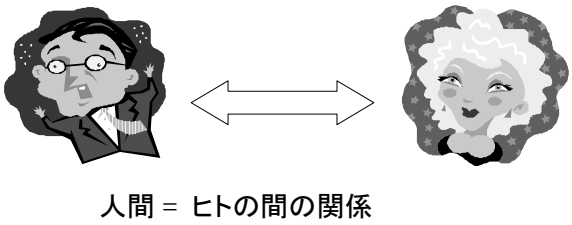


図 3: 人ではなく、人間をモデル化

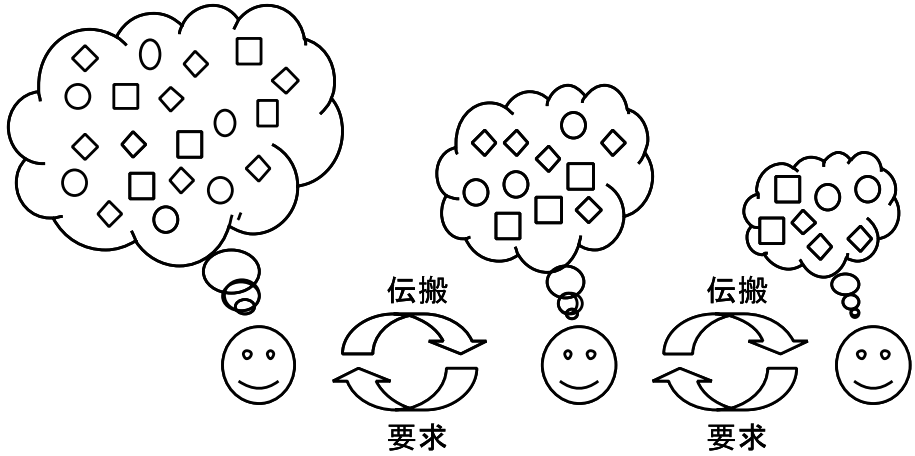


図 4: 口コミのイメージ

多くの人々の評価と伝播を経て、有用な情報だけが人々の間に広まっていく。

口コミによる情報収集では、相手を意識して情報伝達するので、相手の興味のありそうな情報であるかを判断して伝えることや、個人の情報収集では見逃してしまうような重要な情報を伝え合うことができる。さらに、情報を受け取る側は、情報提供者が信頼できる人であるか、あるいは、その話題に関して専門性を持っている人であるかといった評価をあらかじめ持っている。また、本人の経験を元に重要な部分が強調されるので、情報の質が高くなる。これらのことから、情報に対して容易な価値判断ができる。

しかしその反面、「平均化」（伝達要素の減少）や「強調化」（少数要素の強調）、「同化」（予期的枠組みへ一貫する方向への内容変化）などの情報内容の変容が起こりやすく、うわさや流言が発生しやすい。また、人間関係のつながりにより情報が伝播していくので、強いインパクトのある流言でもない限りは、狭い範囲に限定されがちである。

文献 [18] によると、口コミによる情報伝達については、いろいろな研究がされており、その一つに、社会ネットワーク分析[1][25] がある。

親友のように、頻繁に対面接触する緊密な人間関係を「強い紐帯（ちゅうたい）」と呼

[1]集団成員間のコミュニケーションの構造を見出す分析手法。誰と誰がどれくらいコミュニケーションをとったのか、データを採り、これをグラフ化し、グラフ理論を用いて、中心性、密度といった指標を計算することによって、クリークやブリッジといった構造的役割を発見する。

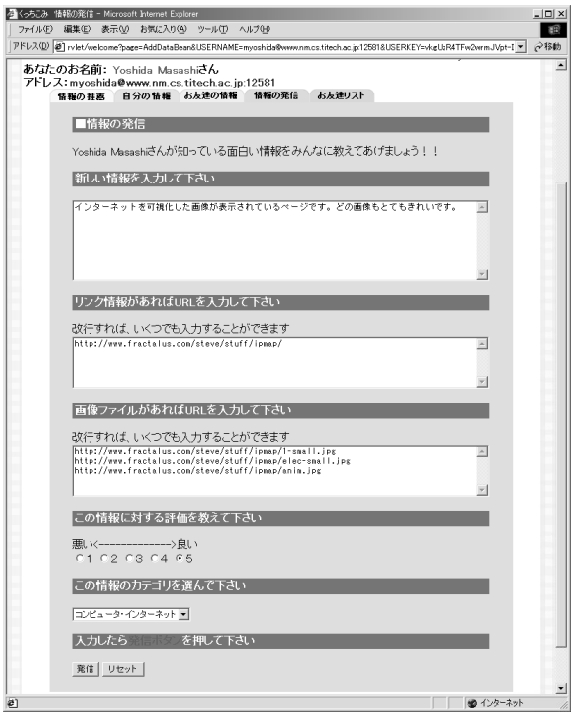


図 5: 情報の発信画面

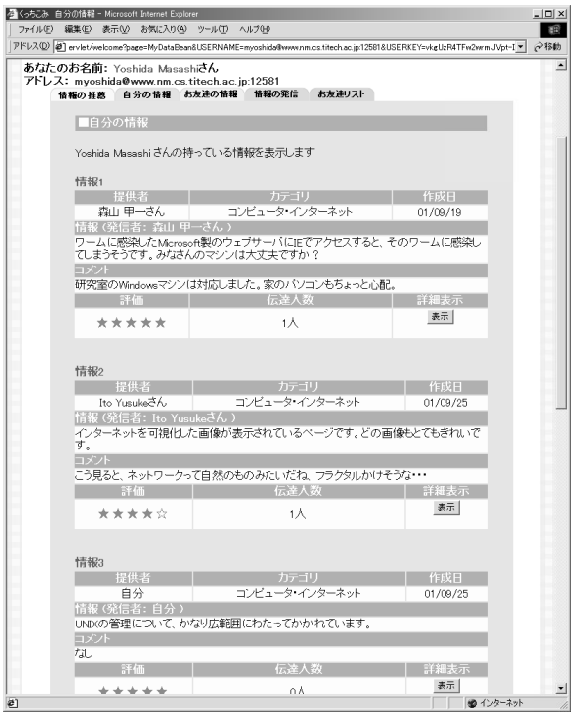


図 6: 自分の情報画面

び、まれにしか対面接触をしない薄い人間関係を「弱い紐帯」と呼ぶ。また、広がる人間関係の網の中で、派閥のように互いに直接結びつきあっている人間関係の集合を「クリーク」という。そして、クリークの間を結びつける弱い人間関係が「ブリッジ」である。

一般に、口コミは、クリークの中で活発に行なわれ、人間関係の紐帯が強いほうが影響力があり、信頼性が高く、有効である。また、専門性を認知されたり信頼性が高い方が、より説得的であるなどの結果も出ている [1]。しかし、転職時に弱い紐帯を通じて情報を得た人の方が転職後の満足度が高く、弱い紐帯は強い紐帯よりも有効であったという興味深い結果も報告されている [4]。これは、クリーク内では情報伝達が早いが、同じような興味が持った人々が集まっているので似たような話題についての情報交換がされやすく、新たな情報はブリッジを通じてクリークに導入されることによるものである。

これらの口コミの諸性質は、電子コミュニティに口コミが再現された場合においても当てはまると考えられる。それと同時に、電子メディア特有の性質により実世界の口コミが持っていた欠点を解消することができる。

電子メディアにおける情報の伝播は、オリジナルのコピーの転送や URL のリンク情報の伝達といった形で行われるため、実世界における情報伝播よりも正確な情報伝播が行われるので、平均化、強調化、同化など情報の変容が起こりにくい [19]。

また口コミは、狭い範囲に限定されがちであるが、電話により口コミの伝達速度が急激に上昇し、その伝達範囲を大きく広げた [18] という事例から考えると、口コミを電子コミュニティ上に再現することで、さらに伝達速度が上昇し、その伝達範囲も広がる。このように、口コミを電子コミュニティ上に再現することは、非常に有効であると考えられる。

4 WAVE

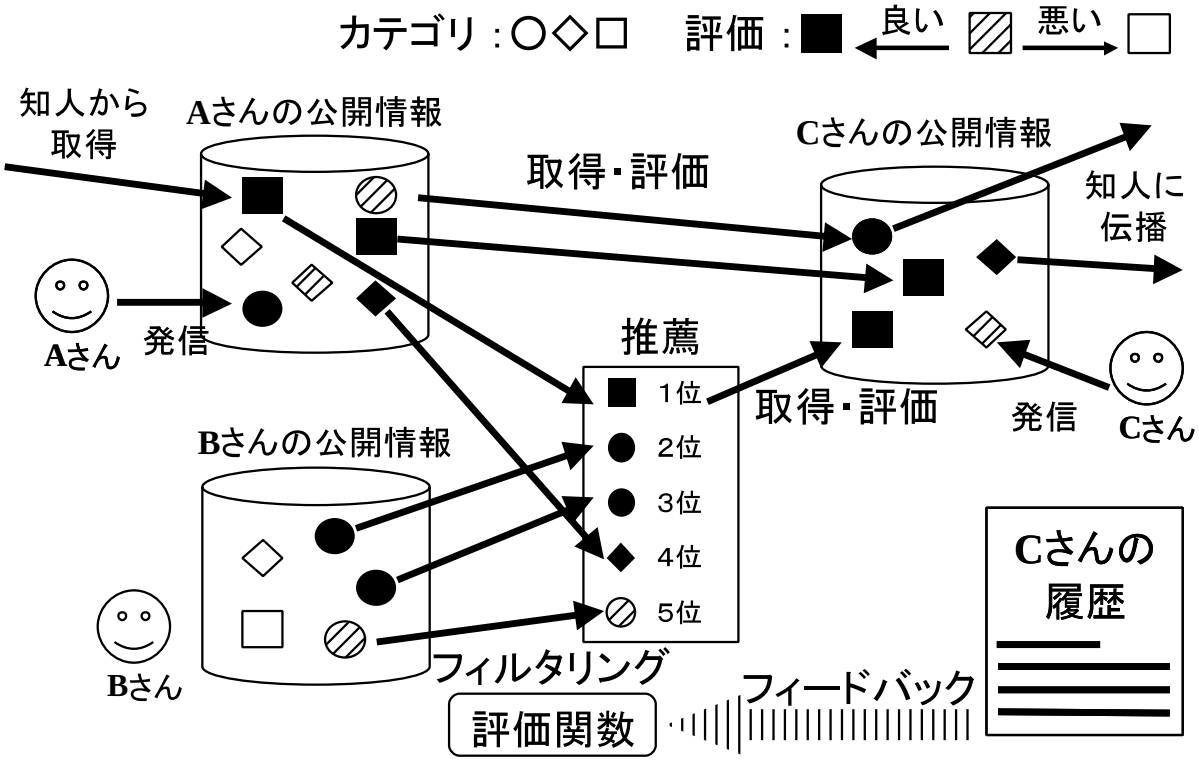


図 7: WAVE における口コミの再現

本稿では、電子コミュニティ上において効率の良い情報収集や円滑なコミュニケーション支援をするシステムとして、WAVE(Word-of-mouth-Assisting Virtual Environment)[2]を提案する。WAVE は、口コミを電子コミュニティ上に再現することで、人間のコミュニティそのものが、分散化された情報収集システムとして機能し、グローバルな情報交換ネットワークを形成する。また、WAVE 上でユーザーが情報の発信、公開、閲覧、評価、取得をシームレスに行うことができるように、ユーザーインターフェースにも工夫を行った。これにより、従来よりも効率のよい情報収集や円滑なコミュニケーションを行うことができる。以下では、WAVE の仕組みとその特徴について図 7 に沿って説明する。

4.1 情報の発信

各ユーザーは、自分が持っている新しい情報を発信することができる。図 5 のように、発信する情報には Web ページや画像データの URL 情報を付加することができたり、情報を閲覧するユーザーが情報の内容を判断しやすいように、情報を簡単に分類するためのカテゴリを割り当てるようになっている。また、情報に対して 1～5（1が一番悪く、5が一番良い）の評価値を与える。発信した情報は、自分の情報として他ユーザーに公開され

[2]WAVE には、口コミが波のように伝播していくというイメージと、このシステムが、WAVE を起こし、世界中の人々に使ってもらえるようなコミュニケーションメディアとなってほしいとの願いがこめられている。

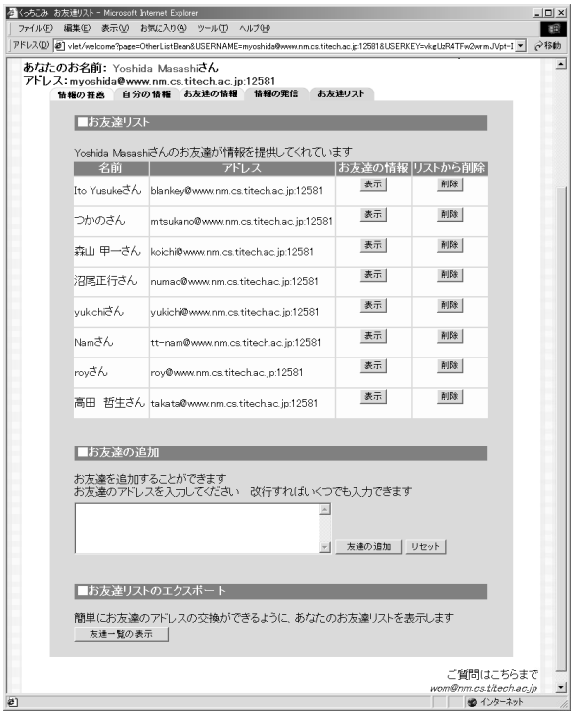


図 8: 情報参照ユーザーのリスト画面

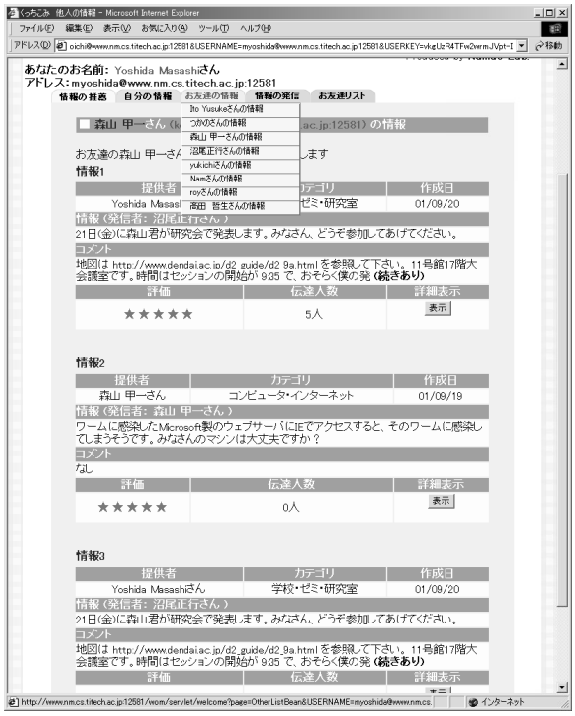


図 9: 他ユーザーの情報画面

る。つまり、WWW やメーリングリストのように、自分なりの情報を多数の人々に向けて自由に発信することが可能である。

4.2 情報の公開

自分が発信した情報や他ユーザーから取得した情報は、自分の情報として公開される。図 6 のように、公開されている自分の情報を閲覧することができ、情報の評価やコメントを修正したり、必要のなくなった情報を削除することもできる。また、図 8 のように、ユーザーの持つ WAVE 専用のアドレス (” ユーザー名@ホスト名:ポート番号” の形) をシステムに登録すれば、そのユーザーの情報を閲覧することができる。

このとき、ユーザーは情報提供者として信頼できるユーザーを登録する。これにより WWW のような情報公開性を持ちながら、電子メールにおける 1 対 1 のコミュニケーションのようにユーザー同士の人間関係が強く現れ、相手の専門分野や信頼性を判断することができるので、WWW などに欠落していた情報源の信頼性を高めることできる。

4.3 情報の評価と取得

図 9 のように、他ユーザーの情報を閲覧する際に興味を持った情報があれば、図 10 のような、その情報に関する詳しいデータを見ることができる。このときユーザーは、その情報に与えられた評価に対して新しい評価をつけたり、新たな付加情報として、コメントを与えることができる。新しい評価やコメントを与えると、その情報は自動的に取得され自分の情報として公開される。そして、さらに別のユーザーによって、その情報は評価・取得されるという繰り返しになる。

図 10: 情報の詳細画面

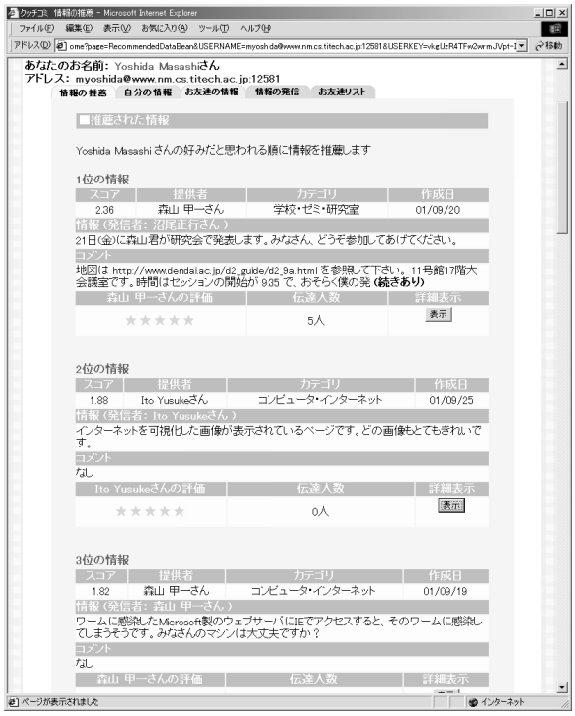


図 11: 情報の推薦画面

つまり、WAVE では情報の公開、閲覧、評価、取得をシームレスに行うことができる。情報をアップロードしたり、情報の存在を人々に知らせたりしなければならないといった、従来 WWW が持っていた情報公開にかかる手間を軽減することができる。また、BBS やメーリングリストにおける ROM(Read Only Member) が持つような、積極的な参加への心理的抵抗感を軽減し、情報伝播において重要な役割を果たすブリッジを維持することができる。

そして、情報の評価と取得が繰り返し行われることにより、人間を介しながら、良い情報だけが生き残り、ネットワーク上に広まる。同じような嗜好を持つ人間は集まりやすいということから、情報の流れは指向性を持ち、ブリッジを介して、情報がひとたび自分の属するクリークに到達すれば、自分もその情報を得られる。つまり、WAVE は、口コミを電子コミュニティ上に再現することで、人間のコミュニティそのものが、分散化された情報収集システムとして機能し、グローバルな情報交換ネットワークを形成する。

カテゴリを情報の発信時に付与し、取得時に変更することができる[3]。これにより、各ユーザ独自の分類が可能である。各ユーザごとに得意分野があるので、他ユーザにとっては、そのユーザの得意なカテゴリのみを参照するのが便利である。

4.4 情報の推薦

他ユーザーの公開している情報を閲覧する際に、情報参照ユーザーの数が増加したり、ひとりのユーザーが公開する情報の数が増加したりすることによって、すべての情報を閲覧するのは、ユーザーにとって負担となってくることが予想される。

[3]さらに拡張して、ユーザごとに階層構造をもった独自の分類体系を設定できるようにする予定である。

そこで、補助的な機能として、図 11 のように、ユーザーの閲覧履歴等をもとに、評価関数を動的に作成し、公開されている情報の中から有用であると思われる情報の一覧を表示する。これにより、ユーザー間での情報のやり取りが支援され、システム上で、より活発な情報交換を行うことができる。

なお、推薦の評価関数は、以下の 2 つの項目について考慮した。

- 情報提供者が与えた情報に対する評価
- ユーザーの嗜好に基づく情報に対する評価

一般に、口コミにおいて、人は他人から聞いた評価を参考にする。WAVE では、ユーザーは、情報に対して 1 〜 5（1が一番悪く、5が一番良い）の評価値を与えており、それが参考になるが、ユーザーはそれをそのまま評価とするわけではなく、情報提供者に対する信頼性や専門性も考慮する。そこで、ある情報の提供者の公開情報をどれくらい閲覧・取得したかや、その情報提供者に対してどういったカテゴリの情報を閲覧・取得したか、回数を記録しておき、ユーザーが他の情報提供者と比べてその情報提供者にどれくらい依存しているかクリック率を計算し、評価関数に用いる。また、このとき、ユーザーの嗜好や他ユーザーとの人間関係、信頼性は、時間が経るにつれて変化していくので、その情報提供者の公開情報を最後に参照してから経過した日数から、最近その情報提供者にどれだけ依存しているかも調べる。

人は、他人から聞いた評価を参考にするだけでなく、自分自身の嗜好も合わせた上で、その情報が有用であるかを判断する。そこで、ユーザーの嗜好によって、その情報に対して、1 〜 5 の評価値でどれくらいの評価を与えるかを予測して、評価関数として計算する。他のカテゴリと比べてどの程度興味があるか、この情報と同じカテゴリの情報に対して与えてきた評価の平均値を計算したり、今までその情報と同じカテゴリの情報をどれくらい閲覧・取得しているか、回数を記録しておき、クリック率も計算する。

一般に多くの人々の間を伝搬してきた情報ほどユーザーは好むので、情報の伝達人数についても考慮し、伝達人数が多いほど評価を高くする。さらに、ユーザーは新しい情報を好むので、情報が公開されてからの日数が経ったものほど情報の評価は下がるようにする。

情報発信時だけではなく、情報取得の際にもカテゴリを付与できるようになっているので、それを考慮するようにすれば、推薦精度が向上する。カテゴリとして階層構造を許すようにすると、階層構造内で何親等になるかという距離を定義することができ、カテゴリ間の類似度が計算できる。それに基づいて他のカテゴリ中の情報への評価を活用すれば、カテゴリに分類したことによるサンプル数減少を補える [22]。

4.5 サーバの分散化

WAVE では、サーバの分散化も行っている。サーバは Java サーブレットを用いて実装されており、図 12 のように Web サーバー上で動作する。ユーザーは、Web ブラウザを介してユーザー登録をしたホストにアクセスするだけで、情報のやり取りをすることができる。

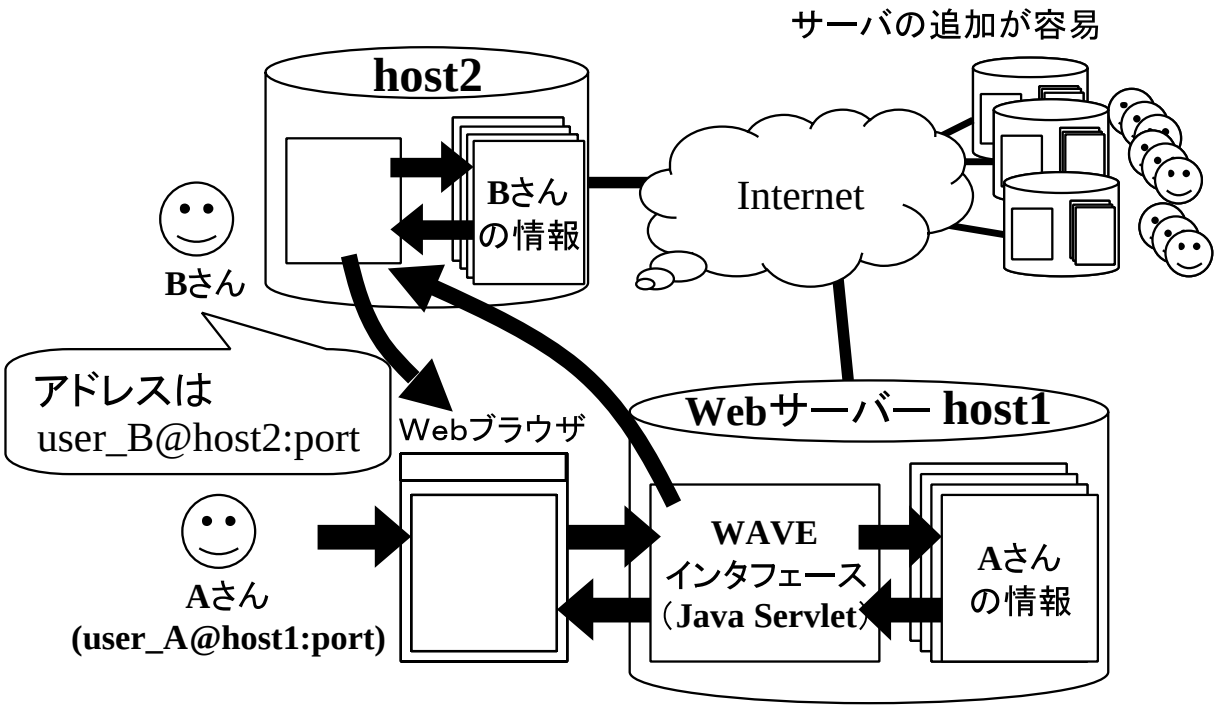


図 12: WAVE サーバの分散化

自分の公開情報は、ユーザー登録したホスト上に蓄積される。なお、この公開情報は、次世代の Web の標準言語となる XML により保存されている。これにより、既存の Web ページからの情報を有効利用することなどが将来的に可能であり、WAVE は WWW を包括するメディアとなり得る。

分散化されたサーバ同士での連携の仕組みは非常に単純で、他サーバーに存在するユーザーの情報の閲覧や取得を行いたい場合であっても、そのユーザーの持つ WAVE 専用のアドレス (" ユーザー名@ホスト名:ポート番号" の形) さえ分かれば、そのユーザーが持つ情報の存在位置が一意に定まるので、ユーザーは、情報の存在位置を気にすることなく情報を獲得することができる。

図 12 を例に説明すると、User1 は、host1 で動作するサーバにユーザー登録しているので、host1 へアクセスすることで、情報のやりとりを行うことができる。また、host2 を利用している友人の User2 の情報が閲覧したいときは、User2 の WAVE 専用のアドレスを登録することで、User2 の情報が host1 に存在するかのようにシームレスに取り扱うことができる。

したがって、分散化によりサーバをどこでも自由に構築することが可能である。さらに、スケラビリティが高く、ユーザー数が増加しても新しくサーバを構築すればよいし、サーバを追加していっても性能はほとんど落ちることがない。また、ネットワークの負荷分散を行うことが可能であり、ＣＰＵ、ストレージなど計算資源を有効に活用できる。さらに、WAVE を Peer-to-Peer 方式[4]で実装すれば、一般的になりつつある個人のインター

[4]クライアント・サーバー型のネットワークと違い、ネットワーク上のコンピュータのそれぞれが、サー

ネットへの常時接続環境において、広く使われる可能性を持っている。このように、分散化によりユーザー数の増加が促進され、より効率的な情報収集を WAVE で行うことが可能になる。

5 一般的な評価

現在、

http://www.nm.cs.titech.ac.jp:12581/wom/

で、システムを公開している。また、本システムを配布し、複数のホストで動作させ、分散環境を実現可能である。興味を持たれた方は是非実験に御協力頂きたい。

本システムを用い、被験者数十人程度の実験を行い、システムの有効性について評価した [15]。情報がどのように流れていくか、情報が伝達するにしたがって評価がどのように変化して行くか、多くの人々を伝達していった情報はどのような内容か、などについてログデータを解析した。また、システムをユーザーに使用してもらった感想を、アンケートを行うことで調査した。そして、ユーザー間で、どれくらい情報のやり取りがなされたか、コミュニティの可視化などを行い、3 章で述べたような、クリーク、ブリッジなどの構造的役割を発見し、システムで口コミが再現されているかを確認することができた。

情報の推薦については、推薦により情報を取得した回数の割合や推薦により取得した情報の順位、アンケートにより、その妥当性を評価した。

6 前処理支援システムとの連携

研究室で開発した前処理用のツール [16] では、データを XML で記述している。WAVE の情報も XML で記述されており、親和性は高い。したがって、前処理過程や結果を WAVE に格納するのは比較的容易である。ここでは、そのツールの概要を述べる。

6.1 前処理

データマイニングではその解析アルゴリズムとして、相関ルール、決定木、クラスタリング、ニューラルネットワーク、遺伝アルゴリズムなど多くのものがある。これらの解析アルゴリズムに大量に蓄積されているデータを適用するためには、何らかの前処理が必要となる。前処理には構造の変形や値の標準化などが含まれるが、これらの作業は事例によって処理が異なり、また経験の求められる複雑な作業であるので、熟練した専門家によって処理される必要がある。そのためにデータマイニングではその処理コストの実に 60% が前処理に費やされている [2]。

現在、前処理の自動化という観点では、属性若しくはレコードの取捨選択を学習によって自動化する研究 [12] や、前処理を行わないまま結果を導出する研究 [8] があるが、現時点で実際に前処理を行う場合は、単純だが有効性が明らかなものを人間であるオペレータが計画を立てて多数組み合わせている。

バーでありクライアントとなる。よって、集中的に処理を行なうサーバーを設置することなく、各ネットワーククライアントが持つ資源をお互いに共有する事が可能である。

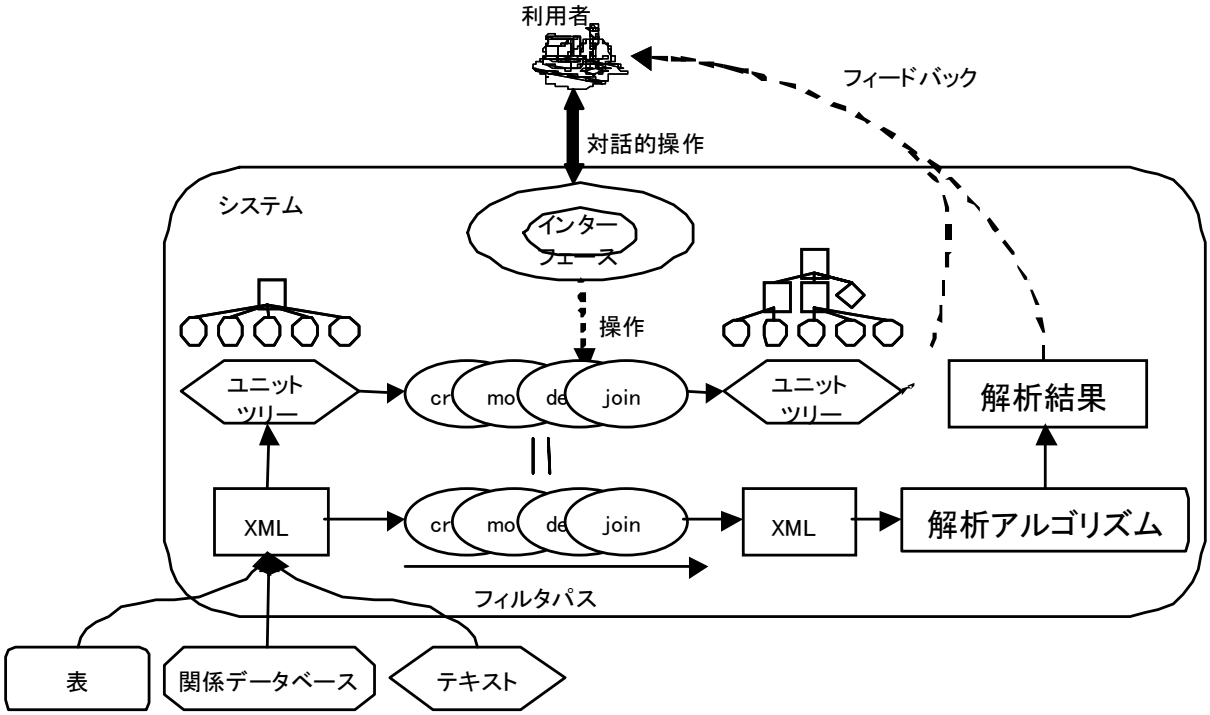


図 13: 前処理支援システム TransX の構成

本研究では、蓄積されたデータに対して前処理を施し、解析アルゴリズムが直接扱えるデータに変換する過程を明確に定義し、その過程に含まれる処理を効率的に行えるデータ構造、及び自動化アルゴリズムの提案と、実際のシステム構築を行った。具体的には、前処理で扱うデータをすべて XML 形式 [3] に統一し、利用者が処理の過程を記述する部分には XML から一意に変換されるデータ構造を用いて効率化を図り、利用者が一度作成したデータ変形フィルタについては自動的に並べ替えて利用者に提案する。

このように、本システム上で扱うデータ形式は、XML 形式に限定している。これは、XML が現在一般的に用いられる表現形式をほとんど表現することのできる表現力を持ち、現在 Web でのデータ交換を中心に標準としての地位を築きつつあるのが理由である。また、広範囲に渡る前処理を行うためには、文字、属性、タプルなど、多くの処理単位を一元的に認識する必要がある。XML は、文脈構造に柔軟性があり、データを処理する上での自由度と処理能力を高めるが、同時に著しい計算時間の増大を引き起こす原因となる。そこで本研究では、データマイニングには全体から見ると小さな構成単位を大量に処理する特色が存在することに着目したデータ構造へ XML を変換し、その変換されたデータ構造に対して処理を行うことで効率化を図っている。実験では、実際の臨床データから決定木を作成する事例について、関係データベースを利用して前処理を行ったものと、本研究で作成したシステム TransX を用いて前処理を行ったものとを比較する。

6.2 従来の前処理の問題点

解析アルゴリズムに対して入力するデータの構造は、一部グラフ構造など他の構造をとる場合もあるがほとんどがフラットな表形式のデータ構造をとる。従って、現時点では前処理に用いるツール、アプリケーションとして表形式のデータ、および表の関係を扱うことのできる関係データベースが用いられる。関係データベースは、大量のデータを高速に処理することができる。しかし、関係データベースを用いた前処理では。表同士の関係の生成や修正、新たな列 (属性) の生成や修正などに大きなコストがかかる。そのため、前処理の最初の段階でデータの構造をしっかり決めておかなければ後の段階でのコストが大きくなる。さらに、実際に前処理を行う上では、バックトラックの管理が必須となる。通常前処理では明確にそのゴールが定まっておらず、前処理を行ったデータを観察したり、解析を行ってみたりなどの作業を行わないと、その前処理への評価が得られないことが多いため、何度も違う方法で前処理を行う必要があるからである。また、同一のデータであっても解析の方法や目的が異なれば前処理もまた異なるものとなる。以上をまとめると、関係データベースよりも強力なデータ構造を持ち、バックトラックの容易性を実現する処理系の実現が必要となる。

6.3 XML とユニットツリー

近年、インターネット上での構造を伴ったデータのやり取りが盛んになってきており、その標準的な形式が XML となりつつある。XML は人間が閲覧、利用する Web 文書とは異なり、計算機によって用いられるものとして設計されており、計算機との親和性が高いといわれている。構造としては木構造をとり、表を組み合わせた構造を取る関係データベースよりも複雑なデータを単純に表現することができる。XML は様々な処理を行うアプリケーションとの親和性についても問題がなく、多くの蓄積された情報が XML に変化する潮流も感じられることを考慮すると、XML を用いて前処理を行うというのは自然な流れである。本研究では主にデータ構造の変形を、XML 変形を用いた処理として実現する。この際に、自動的な前処理が可能となるよう配慮する。つまり、変形の単位を設定し、それをフィルタと呼ぶ。ここでフィルタの大きさを考えた場合、単純に考えると XML の要素を 1 つ作成したり、移動したりすることが 1 つのフィルタであると言えそうだが、これでは要素数の増大により、フィルタの数も増大し、操作が困難となる。そこで、効率的なフィルタを作成するために、XML 全体を一度に把握が可能なユニットツリーと呼ぶ構造を提案する。ユニットツリーは、そもそもデータマイニングにおいてデータ 1 つ 1 つの内容はあまり重要ではなく、全体が表す情報が重要であることに着目し、文書実体から見て同一の階層にある同一の名前を持つ要素を同一とみなす構造である。DTD と似ていると思われるかもしれないが、DTD はそれに適合する XML の集合全体を表しており、特定の XML の状態を表しているわけではない。

6.4 TransX システム

ユニットツリーを用いた前処理支援システム TransX の構成を図 13 に示す。入力された XML ファイルはユニットツリーに変換され、利用者はユニットツリーを見ながら Web ブラウザ上に用意されたインターフェースを用いてフィルタの組み合わせであるフィルタ

パスを構成していく。解析時にはフィルタパスが XML に適用され、その結果を利用者にフィードバックすることができる。利用者はフィルタ単位で前処理を構築することができ、ユニットツリーによってデータの状態を適切に把握することが可能である。

6.5 比較

TransX 上で行った前処理と、既存データベースアプリケーションなどを用いた前処理では、特に処理順序と構造の変更に対する処理について TransX 上で行った方が直感的で扱いやすく、有用である。しかし、XML の持つ潜在的な無駄の多さと実装上の理由から TransX ではまだ膨大なデータを実時間で処理することができない。

6.6 結論

XML 変形を用いたデータマイニングにおける前処理は、有用であり、実用性の高いものに発展可能である。実用性を高めるためには、

- データ量の増加に耐えられる XML 処理系
- XML 定義を完全に満たす XML 解析器の登場が待たれる。

今後の課題としては、XML の特徴を活かした前処理系の提案、変形に対しての XML 問い合わせ言語の使用、解析アルゴリズムを統合した処理系の提案、などがある。

7 アクティブマイニングへの適用

7.1 運用形態

口コミシステムを用いて、全国もしくは全世界に散らばった 1 〜数研究室よりなる研究グループが 10 程度集まり、数十人の人間が連携しながら、共通データをマイニングする場合を考えよう。システムのレスポンスを速くするため、4.5 節で述べた手法により、各研究室ごとに分散してシステムを配置する。ローカルのメンバー間の議論がそこで行われ、他のグループもそれを参照可能にしておく。ただし、外の人間がすべてを読むのは煩わしいので、他のグループでは研究室のゲートキーパー (gatekeeper)[17] の発言のみを参照することになる。このことにより、プロジェクト全体の議論とローカルな議論を区別できる。ゲートキーパは自然発生的に出現するものであり、WAVE はその過程を支援する。

こうしたダイナミックな集団形成を、従来のメーリングリストでサポートしようとすると、複数のリストを使い分けたり、メンバーの入れ替えを頻繁に行ったりする必要が生じ、かなり煩わしいことになる。ニュースシステムでサポートする場合には、ニュースグループの追加/削除を頻繁に行う必要がある。逆に、メンバー全員に周知せねばならない場合には、メーリングリストが有効であり、WAVE はメーリングリストの機能を代替するわけではない。

データを提供する研究室は、口コミシステムで公開し、データについての評判が徐々に流布する形を取る。並行してメールなどで取得を催促することも必要かもしれない。

前処理の方法や前処理結果は、公開したデータにコメントを付け加える形で、追加していく。前処理後のデータについてマイニングを行った結果に応じて評価を与えるようにす

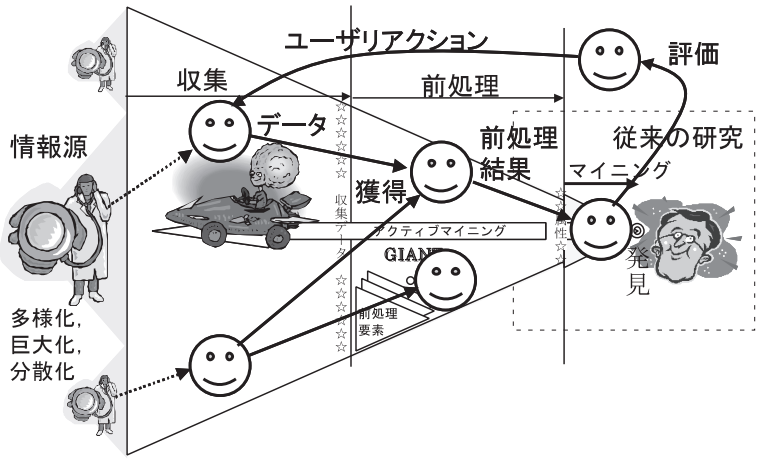


図 14: 口コミとアクティブマイニング

れば、よりよい処理が流布することになる。すなわち、図 14 のように、伝言ゲーム型の口コミ (図 4) で、アクティブマイニング (図 1) を行うことになる。

7.2 情報の推薦方式

各ユーザが 1 〜 5 の評価値を与えているが、さらにマイニング結果の精度 (precision) や再現率 (recall) などを取り入れることも考えられる。推薦は、情報提供者への過去の評価、ユーザの嗜好、経過時間、伝達人数を考慮に入れているが、データマイニングの場合なら、情報提供者への過去の評価、ユーザの嗜好、前処理の質、精度、再現率などを基に推薦することになるであろう。

4.3 節で述べたように取得情報に取得者独自のカテゴリを付与し、たとえば、プロジェクトコアメンバ用カテゴリと各研究室用のカテゴリを分けておくと有効である。参照側はコアメンバであれば、コアメンバ用のカテゴリを読むようになるだろうし、研究室のメンバーなら、研究室用のカテゴリを読むようになる。そのためには、参照する「お友達」だけではなく、その人の取得情報中の特定のカテゴリを指定すればよいし、推薦に用いる評価関数にカテゴリを含めてもよい。このような絞込みを行うことにより、余計な情報は目に触れにくくなり、情報の錯綜を最小限に押さえられる。

取得情報にカテゴリを付与できない場合には、仕事に集中するため、データマイニング用のシステムは他とは分けた方がよさそうである[5]。カテゴリが付与できるのであれば、「データマイニング」のようなカテゴリを用意し、関係するカテゴリをそのサブカテゴリとして収容すれば、他の話題との情報錯綜は避けられる。

7.3 セキュリティ

データマイニングの場合、情報の公開範囲はプロジェクト構成員のみに制限する必要がある。これは、通信相手を制限することで可能である。プロジェクトの中ではアクセス制限は設けないが、余計な情報を見なくて済むように、各人の判断で、カテゴリやお友達リ

[5]前述の URL を覗き始めると、しばし脱線してしまう :-)。

ストで制限を加えると同時に、推薦精度を上げることで対処する。

7.4 巨大データの扱い

システムは、情報を取得する際に、過去のコメント全体をやりとりするようになっている。この形でデータマイニング対象の大きなデータ全体をやりとりするのは現実的ではない。そこで、システム内に別ファイルとして置き、メッセージにリンクを添付する形で渡すようにしている。

7.5 帰納論理プログラミングのための新手法開発

グラフリダクションによる推論 [6] において、各アークへ重み付けを行うことにより、情報収集および音楽感性の獲得を行った [7]。この手法は、重み付けにより一階述語論理に相当するネットワーク表現を学習するもので、あいまいな結論を獲得できることに特徴があり、帰納論理プログラミングの頑健性を改善する手法として、データマイニングにおいて有効であると考えられる。また、前処理支援システムと口コミシステムを連携させる際にも利用できる。Prolog 版、Lisp 版、C 版の三つの版をインプリメントしているが、現在、Prolog 版を利用して重み付けの手法を再検討しており、システム全体の実現は来年度の課題になっている。

8 おわりに

本稿では、口コミを電子コミュニティ上に再現することで、効率のよい情報収集や円滑なコミュニケーションを支援するシステムを提案した。それをデータマイニングにおける協同作業に適用する場合についても考察した。

関連図書

[1] J. Bristor. Exhanced explanations of word of mouth communications;the power of relations. *Research in Consumer Behavior*, Vol. 4, pp. 51–83, 1990.

[2] Peter Cabena and Pablo Hadjinian. *Discovering Data Mining*. Prentice Hall PTR, 1998.

[3] World Wide Web Consotium. External markup language (XML). http://www.w3.org/XML/.

[4] M. Granobetter. *Getting A Job*. 1974.

[5] Pattie Maes. Agents that reduce work and information. *CACM*, Vol. 37, No. 7, pp. 30–40, 1994.

[6] M. Numao, S. Morita, and K. Karaki. A learning mechanism for logic programs using dynamically shared substructures. In *Machine Intelligence 15*, pp. 268–284. Oxford University Press, 1999.

[7] Masayuki Numao, Daishi Kato, and Masaru Yokoyama. Learning organization in global intelligence. In *AAAI Spring Symposia*. AAAI Press, 2002.

[8] A. Ragel and B. Cremilleux. Treatment of missing values for association rules. In *PAKDD*, pp. 258–270, 1998.

[9] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architechture for collaborative filtering of netnews. In *CSCW '94*, pp. 175–186, 1994.

[10] S.Brin and L.Page. Anatomy of large-scale hypertextual web search engine. In *Proc. 7th International World Wide Web Conference*, 1998.

[11] S. Wrobel. Scalability, search, and sampling: From smart algorithms to active discovery. In *Lecture Notes in Artificial Intelligence*, Vol. 1268, p. 507, 2001.

[12] Xindong Wu. Induction as pre-processing. In *PAKDD*, pp. 114–122, 1999.

[13] 伊藤雄介, 吉田匡史, 沼尾正行. 口コミ支援システムの実験. 情報処理学会知能と複雑系研究会, Vol. 01-ICS-124, pp. 9–16, 2001.

[14] 伊藤雄介, 吉田匡史, 沼尾正行. 多くの人の評価を経て情報が吟味される口コミ支援システム. 人工知能学会全国大会（第 15 回）論文集, 2001.

[15] 吉田匡志, 伊藤雄介, 沼尾正行. 口コミによる分散型情報収集システム — WAVE を起こそう — Word-of-mouth-Assisting Virtual Environment. In *MACC2001*, 2001. http://www-kasm.nii.ac.jp/macc2001-proceedings/MACC2001-10.pdf.

[16] 五十嵐建平, 大田佳宏, 横山茂樹, 沼尾正行. データマイニングにおける XML を用いたデータ構造の変形. 人工知能学会全国大会（第 15 回）論文集, 2001.

[17] 後藤滋樹, 野島久雄. 人間社会の情報流通における三段構造の分析. 人工知能学会誌, Vol. 8, No. 3, pp. 348–356, 1993.

[18] 中村功. 現代のエスプリ別冊 特集「流行 ••• ファッション」流行と口コミと電話.

[19] 柴内康文. 電子メディア社会における情報伝播. 第 2 回 CMCC 研究会シンポジウム, 1999.

[20] 溝口文雄, 大和田勇人. 帰納学習に基づく情報フィルタリング. 人工知能学会全国大会（第 10 回）論文集, 1996.

[21] 舘村純一. 協調型情報探索を支援する仮想評者とその視覚化. 1999.

[22] 沼尾正行, 横山甲. 階層化された知識の継承による情報フィルタリング. 情報処理学会知能と複雑系研究会, Vol. 99-ICS-116, pp. 43–48, 1999.

[23] 沼尾正行, 吉田匡志, 伊藤雄介. 口コミに基づく情報収集とデータ前処理. 人工知能学会第 46 回人工知能基礎論/第 54 回知識ベースシステム研究会, 第 SIG-FAI/KBS-J 巻, pp. 47–54, 2001.

[24] 高橋正道, 北山聡, 金子郁容. ネットワーク・コミュニティにおける組織アウェアネスの計量と可視化. 情報処理学会論文誌, Vol. 40, No. 11, pp. 3988–3999, 1999.

[25] 安田雪. ネットワーク分析. 新曜社, 1999.

[26] 岡堂哲雄. 現代のエスプリ別冊 社会心理用語事典. 至文堂.

[27] 西田豊明, 畦地真太郎, 藤原伸彦, 角薫, 福原知宏, 矢野博之, 平田高志, 久保田秀和. パブリック・オピニオン・チャンネル. 第 2 回 CMCC 研究会シンポジウム, 1999.

[28] 藤田邦彦, 亀井剛次, Eva Jettmar, 吉田仙, 桑原和宏. ネットワークコミュニティの可能性–community organizer 評価実験結果報告–. 第 3 回 CMCC 研究会シンポジウム, 2000.

多段階学習によるデータ収集と前処理の自動化

研究分担者　櫻井 成一朗　(東京工業大学大学院情報理工学研究科)

HTML のリンク構造と構文的特徴に基づく知識獲得について

背景と目的

WWW の巨大化によって有用な情報の抽出が日に日に困難になってきており，この困難を克服すべく様々な研究が活発に行われている．しかしながら，情報洪水の問題だけでなく，ユーザの意図を的確に把握することも困難であるので，検索エンジンやディレクトリサービスによって情報を引き出すしか方法が提供されていないのが現状である．ユーザの意図を把握するのであれば，ユーザの意図をモデル化した知識を予め与えておくか，ユーザの意図を動的にモデル化できなければならない．ユーザの意図を的確に把握することはもちろん困難であるが，ポータルサイトのように訪問するユーザの意図を最大限に考慮して作られた WWW ページも少なくない．本研究の目的は，ユーザの意図を考慮して作成された WWW ページを手本にして，ユーザの意図をモデル化するための知識を WWW から直接獲得する方法を実現することにある．

WWW を構成する HTML ファイルは半構造化されたハイパーテキストである．ハイパーテキストの特徴を利用した知識獲得という観点では，特定のトピックに関して有用なページ集合を求める方法として，Web コミュニティ発見手法 [1, 2] や HITS [4, 5, 6] がある．Web コミュニティ発見手法や HITS では Web の持つグラフ構造に基づいて関連トピックの知識を抽出する試みである．これらの研究では，ページ作成者がリンクを張ったという行為を特定 URL の推薦行為であるとみなして，関連ページ群を獲得している．換言すれば，これらの方法では作成者の推薦意図を汲み取ってはいるものの，作成者の他の意図を捨ててしまっているのである．ページ作成者がユーザの意図を十分考慮して作成したのであれば，捨ててしまった情報の中にも作成者が考慮したユーザの意図が隠されている可能性がある．本研究の第一の課題は，リンク構造の中に隠された，ページ作成者の作成意図を抽出することでユーザの意図をモデル化する知識を抽出することである．

リンク構造ではなく，構文的特徴を利用して知識抽出を行う研究も数多くなされている．例えば，[3] は HTML から情報を抽出するためのラッパを自動生成する方法を提案し，[10] では定型書式で記述された HTML 文書から XML に変換する方法を提案しているが，いずれも構文的制限が強い．WWW 作成者の多様性を考慮すれば，何らかの構文的制約を課する必要性は否めないが，作成者の意図を汲み取るためには，より単純な構文的制約でも十分な場合が少なくない．本研究の第二の課題は，より単純な構文的制約の下で，作成者の意図を抽出し，ユーザ意図のモデル化に供することである．

本研究では，WWW ページを検索するための知識として分類知識と属性名と値の組からなる集合として表現される概念知識を対象とする．これらの知識を獲得するために，一定の構文的特徴を利用して WWW ページを絞り込み，知識獲得する方法を提案する．

検討内容

知識獲得の対象

本研究では，知識獲得の対象を一定の構文的特徴を有する WWW ページとする．２０億ページにも及ぶような WWW ページでは様々な書式でデザインが行われ，また純粋な文書としても様々な表現が採用されているため，網羅的にすべての WWW ページから知識獲得を試みることは効率的ではないからである．構文的特徴を持つ WWW ページとしては，リンク集とデータ集を考える．リンク集とは，複数の WWW ページへのリンクを集めたページの事である．各リンク集は作成者の利便性向上のため，あるいは他の閲覧者に有用なページを提示するためなどの目的で作られており，その目的が検索者の意図と一致する事が少なくないと考えられるからである．データ集とは，属性名と属性値の組の集合を記述しているページの事である．データ集は個人の自己満足のためだけに作成されたものもあるが，他人に閲覧されることを前提として作成されたものが多いからである．他人に閲覧されることを想定して作成するのであれば，どのような意図で当該ページを閲覧するのかを熟慮して作成されたページである事が期待できる．

基本アイデア

本研究における基本アイデアは，現在の検索エンジンのテキスト照合の高速性を利用することで，構文的な特徴を共有する WWW ページを収集することにある．リンク集についてはバックリンクページがリンク集である頻度が高い事を利用し，データ集については予め XML の雛型を与えておく事によって検索エンジンを利用する．収集した WWW ページから知識を獲得するには，HTML 構文に関する制約を利用して知識を抽出する．

リンク集の収集

膨大な WWW ページの中からリンク集だけを収集する事は容易ではないと想像されるが，特定の URL にリンクを持つバックリンクページの多くがリンク集であることが予備実験 [12] の結果確かめられた．例えば，トヨタ，ホンダ，日産の WWW ページへのリンクを同時に持つページは，自動車会社リンク集である可能性が高くなるということが [1, 2] の基礎をなしている．実際に「トヨタ」と「日産」のバックリンクページをアルタビスタ (http://www.altavista.com) によって調べた４１９ページについて各バックリンクページの持つアンカータグをヒストグラムとして示したグラフが図 1 である．図１ではバックリンクページの半数以上が５０以上のアンカータグを持ち，アンカータグの平均は１６９であった．予備実験では，他の分野の URL も同様に選んでみたが，自動車会社と同様にリンク集が多いことが確かめられた．

リンク集の構文的特徴

リンク集は，作成者個人のためと言うよりも，他の利用者のために作成されていることが多い．他の利用者に使ってもらうためには，視覚的な要素に訴えるということが必要であり，視覚的に工夫されたデザインが実際に少なくなかった．しかしながら，画面配置を工夫するには限定的な機能しか持たない HTML を用いるのであれば，利用できる言語機能は自ずと限られてくる．もちろん，マルチメディア拡張機能を利用することも考えら

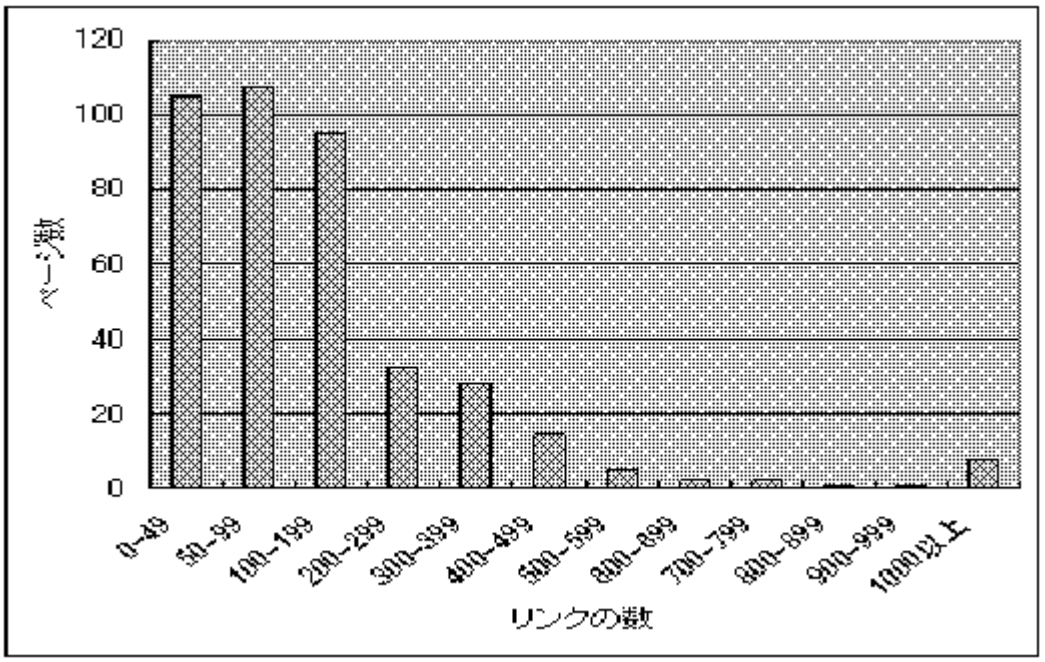


図 1: **リンク集としてのバックリンクページ** [12]

れるが，リンクの数が増大すると，マルチメディア機能によるアクセス負荷が増大してしまっては逆効果になってしまう．したがって，リンク集は一定の構文的特徴を有することになる．すなわち，ページ作成者は URL **に対して何らかのカテゴリを設け，そのカテゴリ分類にしたがって** HTML **デザインを行っていると考えられるのである．そのカテゴリを「見出し」として提示し，ページ利用者を助けているのである．この作成者の分類知識は，検索のための重要な知識源となる事が期待されるので，作成者の分類知識の獲得方法について考察する．**

各リンクを画面上で整列させるためには，リスト形式の構文やテーブル（表組み）の構文が多用されることになる．実際のリンク集で採用されている構文の多くは，図 2 **に示すリストを用いるものとテーブルを用いるものが多かった．図** 2 **の例では，リンク先の分野を表すテキストの後にその分野のリンクのリストが並んでいる．カテゴリ毎に異なるリストになっており，**HTML **を構文解析することで分割できる．テーブルの場合には，分野名を表すテキストの後にリンクが入っているテーブルが続く，という構造が連続している．**HTML **を構文解析することによってリストの場合もテーブルの場合も分野毎に分割する．**

データ集の収集

データ集から獲得する知識は図 3 **に示す** XML **によって表現するが，属性名と値の対集合によって概念を表す．属性名については** XML **の** DTD **として予め与えておくものとする．図** 3 **では，**

$$\{< \textbf{タイトル}, \textbf{タイタニック} >, < \textbf{監督}, \textbf{ジェームス・キャメロン} >, ...\}$$

```
<h3> Car Manufacturers On the Web</h3>
<ul>
  <li><a href="...">...</a>...</li>
  <li><a href="...">...</a>...</li>
  ...
</ul>
<h3> Favorite Jetskies</h3>
<ul>
  <li><a href="...">...</a>...</li>
  <li><a href="...">...</a>...</li>
  ...
</ul>
```

(a) **リストによるリンク集**

```
<table>
  <tr>
    <td>
     Car Manufacturers On the Web
    </td>
    <td>
      Favorite Jetskies
    </td>
  </tr>
  <tr>
    <td><a href="...">...</a>...</td>
    <td><a href="...">...</a>...</td>
  </tr>
  <tr>
    <td><a href="...">...</a>...</td>
    <td><a href="...">...</a>...</td>
  </tr>
</table>
```

(b) **テーブルによるリンク集**

図 2: **リンク集の例** [12]

<映画>
 <タイトル>タイタニック</タイトル>
 <監督>ジェームス・キャメロン</監督>
 <出演>レオナルド・ディカプリオ</出演>
 ...
</映画>

図 3: 獲得すべき XML の例 [11]

という集合として「映画タイタニック」の知識を表現している．データ集の収集 [11] に関しては，XML の属性タグによって検索エンジンを呼び出すが，属性タグを連言として与えるので，データ集となることが期待できる．

データ集の構文的特徴

データ集も，他の利用者に使ってもらうためには，視覚的な要素に訴えるということが必要であり，視覚的に工夫されたデザインが実際に少なくなかった．したがって，データ集も一定の構文的特徴を有することになる．すなわち，統一的な見栄えを実現するために，定型句の繰り返しが多くなるのである．

[10] のような強い構文上の制約を緩和するために，単純なテキスト照合のみで値を抽出する方法について考察し，データ集からの知識獲得システムを実装した．対象としては，単一の概念を記述したページではなく，複数の概念を大量に記述したページを選択した．単一の概念しか記述していない場合には，ページ作成者が多様であるので構文的規則性を抽出する事が困難であるが，同一ページ内で複数の対象を記述するものから構文的規則性の抽出を試みた．

まず訓練例を予め用意しておき，属性と値の間の文字列を文字列パタンとして抽出し，文字列パタンを使って HTML ファイルから XML を獲得した結果を表 2 に示す．文字列パタンを適用する HTML ファイルの収集時には，属性名に関しては予め類義語を用意しておき，類義語を含む HTML ファイルも対象に含めた．文字列パタンは属性名と値の部分が変項であり，変項の間の文字列は固定項である．文字列パタンは，構文木のような複雑なラッパとは異なり，構文解析せずに単なるテキスト照合で済むので，高速に値を抽出できる．もちろん，文字列パタンが汎用的に有効というわけではなく，データ集を記述する際には，HTML 構文が限定されてしまうという理由によるものと考えられる．この理由はリンク集においても特定の HTML 構文が多用されてしまう事と一致している．

結果

リンク集からの知識獲得システム

入力された URL からリンク集の分類知識を獲得するシステムを CGI プログラムとして実装した．この CGI は一つまたは複数の URL に対するバックリンクからリンク集を

収集し，構文的特徴によってリンク集を分割する．尚，構文解析にはPerlライブラリのHTML-Parserを利用した．

リンク集のHTMLファイルの構文解析に成功すれば，その構文解析木から構文的特徴に基づいてURL群を抽出することができる．リスト構造に関してはリストの各要素についてアンカーが連続して出現しているものを一分割とし，アンカーを含まない要素については項目の区切りであると見なして分割する．テーブルによる2次元配置では，列毎の分割と行毎の分割を考慮した．行と列のどちらの分割を採用するかは，アンカーを含まない項目の出現位置によって判断することとした．トヨタと日産を入力として獲得したリンク集を表1に示す．

グループ名	上位5個のURL
新聞社	http://www.nikkei.co.jp/, http://www.asahi.com/, http://www.yomiuri.co.jp/, http://www.mainichi.co.jp/, http://www.sankei.co.jp/
検索エンジン	http://www.yahoo.co.jp/, http://www.goo.ne.jp/, http://www.infoseek.com/, http://www.yahoo.com/, http://www.lycos.co.jp/
コンピュータ関連会社	http://www.melco.co.jp/, http://www.sony.co.jp/, http://www.nec.co.jp/, http://www.ibm.co.jp/, http://www.toshiba.co.jp/
旅行関連会社	http://www.jtb.co.jp/, http://www.jal.co.jp/, http://www.knt.co.jp/, http://www.nec.co.jp/, http://www.fujibank.co.jp/
放送局	http://www.tv-tokyo.co.jp/, http://www.nhk.or.jp/, http://www.ntv.co.jp/, http://www.tv-asahi.co.jp/, http://www.tbs.co.jp/index-j.html
自動車関連組織	http://www.jaf.or.jp/, http://www.osa.go.jp/, http://www.motorshow.or.jp/, http://www.j-sapa.or.jp/, http://oil-info.ieej.or.jp/

表1: 収集できた周辺URL群 [12]

表1に示したように，自動車会社以外のURL群として，新聞社，検索エンジン，コンピュータ関連会社，TV局のURL群が得られた．各グループ名は人手によって付けたものであるが，各群の要素は自動的に選択されたものである．

データ集からの知識獲得システム

文字列パタンの獲得システムと文字列パタンを利用してHTMLファイルからXMLを獲得するシステムを実装した．表2に示した結果から，WWWの多様性を考慮すれば比較的良好な結果を得る事ができたと考えられる．

XML獲得の失敗例の中では属性名に相当するキーワードが欠落しているHTMLファイ

	対象 URL	獲得した XML
映画	706	594 (84.1%)
温泉	180	108 (60.0%)
本	296	233 (78.7%)
合計	1182	935 (79.1%)

表 2: XML の獲得数 [11]

ルが顕著であったので，主たる属性については，属性名が欠落している場合でも検索できるようにシステムの拡張を行なった．すなわち，タグ補完機能を実現することで，表 3に示すように 7 割程度の XML が獲得できるようになった．

	獲得可能な XML(HTML)	補完に成功した XML
映画	1207(27)	831 (68.9%)
温泉	57(11)	48 (84.2%)
本	105(23)	95 (90.5%)
合計	1369(61)	974 (71.2%)

表 3: タグ補完の効果 [11]

考察

リンク集からの知識獲得

多数のページから参照される大企業の URL を選ぶ事で，従来の Web コミュニティ手法とは異なり，複数の URL 群を比較的容易に獲得することができた．獲得した URL 群については，トピックグループとして直接利用することも可能であるが，次段の学習における入力として利用することが考えられる．そのためには，リンク集から獲得した知識も XML として蓄積しておくことが考えられる．

獲得した URL 群の洗練化については，直接 Web コミュニティ手法を用いることも可能であり，実際に Web コミュニティ手法の適用も試みたが，まだまだ改善の余地がある．実際に表 1 の旅行関連会社の中には関係が薄いと思われる www.nec.co.jp が含まれていた．このような分類誤りを解消することも今後の課題の一つであるが，より重要な課題として獲得したグループの名前付けの問題がある．表 1 の場合には，人手でグループ名を付与したが，グループ名の与え方については来年度以降検討したい．

データ集からの知識獲得

XML として獲得すべき知識の雛型を与える事が対象となる HTML ファイルを制約するので，単純なパタン照合でも比較的良好な結果を得る事ができた．獲得された XML については，XQL のような XML 用検索エンジンを利用することで，WWW を検索する代

わりに利用する事ができる．例えば，映画の XML を構築できれば，出演者から映画の題名を検索したり，題名から監督を検索できる．

データ集から獲得した知識の２次利用法としては，やはり次段の学習の入力として利用する事が考えられる．具体的には，特定の概念の必須属性や非必須属性を決定するのに利用できる．

前処理支援のための知識獲得

本研究で獲得した知識は，直接知識処理に適用することもできるが，更に加工すべき知識でもある．具体的には，リンク集から獲得した知識に関してはなぜ作成者がそのような分類を行ったのかについては考慮しておらず，データ集から獲得した知識に関してはなぜ作成者がその属性を選択したのかについて考慮していない．分類とは，目的志向によるものであり，目的が違えば自ずと異なる分類を採用しなければならない．属性の選択も同様である．お手本となる WWW ページから作成者の意図を汲み取ることで，ユーザの意図を汲み取るのであれば，これらの作成者の考える理由付けについて考慮しなければならない．この点に関しては，仮説推論や ILP 手法を適用することで，作成者の意図を関係的知識として汲み取ることを現在検討している．

前処理においては，ユーザの負担を軽減することはもちろんであるが，ユーザが理解しやすい形で情報提示する事が重要であり，情報提示の方法についても現在検討中である．

今後の課題

今後の課題を以下にまとめておく．

- リンク集から獲得したグループに対する名前付け
- 獲得した知識を利用した検索エンジンとのインタフェース作成
- リンク集から獲得した知識とデータ集から獲得した知識の統合
- 仮説推論や ILP を用いたページ作成者の意図の抽出

参考文献

[1] 村田剛志：“参照の共起性に基づく Web コミュニティの発見”，人工知能学会誌，Vol. 16，No. 3，pp. 316-323(2001).

[2] 村田剛志：“Web コミュニティにおける構造モデル”，情報処理学会研究会，予稿 (2001-ICS-124)，pp. 41-45 (2001).

[3] W. W. Cohen:“Recognizing Structure in Web Page using Similarity Queries”，Proc. of AAAI-99，pp. 59-66 (1999).

[4] Jon M. Kleinberg: “Authoritative Source in Hyper linked Enviroment”, Proc. ACM-SIAM Symp. on Discrete Algorithms, pp. 668-677(1998)

[5] Soumen Chakrabarti, et. al.: “Mining the Web’s Link Structure” ,IEEE Computer, vol. 32, no. 8 (1999)

[6] Soumen Chakrabarti, et. al.: “Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks”, SIGIR 2001: http://www.cs.berkeley.edu/∼soumen/

[7] **廣川佐千男, 池田大輔：**“Web **グラフの構造解析**”，**人工知能学会誌，**Vol. 16，No.4，pp. 525-529(2001).

[8] Page, L. Brin,S. Motwani R. and Winograd T.: The Page Rank Citation Ranking: Bringing Order to the Web，Online manuscript，http://www-db.stanford.edu/∼backrub/pageranksub.ps (1998).

[9] Kumar,R. et al.:”Trowling the Web for Emerging Cyber-Communities”, Proc. of the 8th WWW Conf. (1999).

[10] **梅原雅之，岩沼宏治：**“**事例に基づく** HTML **文書からの** XML **文書への半自動変換**”，**人工知能学会誌，**Vol. 16，No. 3，pp. 316-323(2001).

[11] **山田作：**“**タグの共起性に基づく** HTML **文書からの** XML **獲得に関する研究**”，**東京工業大学大学院情報理工学研究科　修士論文** (2002)

[12] **若月謙太郎：**“**リンク解析による** WWW **からの知識獲得に関する研究**”，**東京工業大学大学院情報理工学研究科　修士論文** (2002)

発表論文

[1] **若月謙太郎，櫻井　成一朗:リンク解析による** WWW **ページ群の発見，人工知能学会，予稿** (2001-FAI-46, 2001-KBS-54)，pp. 109-114 (2001).

高次元インデックス技術を用いた検索処理性能向上について

研究分担者　河野 浩之　(京都大学大学院情報学研究科)

1　研究背景

地理情報システム (GIS:Geographic Information System) を用いたシステム構築，地理情報データを交換するための空間データ整備が活発化している．そして，交通管制・交通計画・都市計画などへの高度活用を目指したアプリケーションとして，道路網上に配備された各種情報センサーから取得した様々なデータを蓄積する交通データウェアハウス (Data Warehouse) がある [7]．実際，情報センサーの高度化により，広範な領域から，車両位置をリアルタイムに高精度収集することが検討されている [8].

そこで，我々は，位置データに対する検索処理の基礎技術である，データベース研究領域における空間データベース，あるいは，時空間データベース技術に注意を払っている [3]．特に，空間オブジェクト検索や解析性能に多大な影響を及ぼす空間インデックスや時空間インデックスに関する研究に重きを置く．また，データマイニングの領域における，空間データに対する空間データマイニング (spatial data mining) 研究 [12, 4, 15] の適用可能性にも注目してきた.

なお，我々は，この種の高次元データ活用を目標に，検索及び解析処理性能の向上を目指した研究を行ってきた [3]．すなわち，R-Tree,　R*-Tree, PR-Quadtree などの空間インデックス技術 [17] を利用した空間データマイニング [9, 10]，TPR-Tree, 3DR-Tree などの時空間インデックス技術を用いた OLAP (On-Line Analytical Processing) などである.

ところで，我々の具体的な目標は，地理情報システムの戦略的活用であり，都市計画，道路計画，最適経路，施設の最適配置等，社会生活において重要な位置を占める多くの現実問題 [14] に少なくとも対応できることである．すなわち，個人の移動経路に付随する各種データの記録・蓄積を行うことで，交通行動調査 [2]，パーソントリップ調査 [1] に新たな展開を与えることを目的とする．また，交通管理や交通管制において，位置情報システムと GIS を効果的に連携させるための検討を行っている [20]．なお，この種のシステム構築技術を確立することで，現在使用されている定点型車両検知器から得られるデータに基づいた交通管制・情報提供に関わる交通流解析の問題点の改善が可能である．このように，交通管制，交通計画，マーケティング等における意思決定支援システムに実装可能なことを目標に研究を進めている.

そこで，この種のデータ解析を効果的に実行できるシステム構築技術を確立するため，空間インデックス技術を利用した位置データ処理精度の向上，蓄積された空間データに対する経路推定 [11, 2]，さらに，交通情報システムにおいて必要とされる各種空間問合せ [4] に関する検討を進めてきた．また，移動経路推定，定時刻問合せ，時間帯問合せ，等速移動領域問合せを対象に，大阪市周辺における実測データを用いて，そのシステム性能に影響する各種パラメータを考察し，システムアーキテクチャの検討を行ってきた.

以下，2 章では，位置情報システム構築を行うにあたって生じる幾つかの問題を示す．3 章では，本研究を遂行する上で，主要な役割を果たす空間インデックスや時空間インデッ

クスなどの高次元インデックス技術を紹介し，交通工学領域における典型的な問合せと関連付けて述べる．そして，今後，研究を推進する方向を 4 章で述べる．

2　位置情報システム構築に関わる問題

本章は，PHS などから得られる位置データを，交通管制・情報提供やパーソントリップデータ調査に利用するシステム構成を検討する上で障害となった幾つかの問題を紹介する．なぜなら，現実社会における問題を対象とするからであり，理想的条件を仮定した技術的議論だけでは，実システム構築は不可能であるからである．そして，この問題は，実データを処理対象とするデータマイニングにつきまとう問題でもある．

【地理情報システムに関わる問題】

地理情報システムは，道路や建物等の空間属性と，道路名称や建物所有者のような非空間属性を相互に関連付けたデータベースを用いて，検索・解析・表示処理などを行うものである．しかし，多くの地理情報システムは，目的別に設計されており，精度の異なる都市情報や道路情報を蓄積したものである．そのため，本研究の目標となるような精度の高い交通データウェアハウスを構築するに，現在の地理情報システムを利用することは非常に困難である．例えば，複数の都市交通機関，地下街など，管轄の異なる機関によって作成された地理情報を統合するための技術的提案があっても，実用化されておらず，また，データ流通そのものが十分でないからである．そのため，実験を進めるにあたって，基礎データである地図データ整備から必要となった．

【位置データ処理に関わる問題】

我々の目的に合う精度をもつ地理データ整備のために，国土地理院発行の 1/2500 レベルの解像度をもつ数値地図 2500 を用いた [20]．まず，回転楕円体である地球の位置を「平面直角座標系」で表す変換処理や，5 系と 6 系で表記される近畿地方の地図接続のための座標変換処理が必要となった．加えて，空間データ利用の用途に応じて表現形式が異なるため，位置情報システムと連携するには，WGS-84 系，ITRF(International Terrestrial Reference Frame) 系などの変換も要した．

また，位置データ獲得のために利用可能な各種デバイスの平均測定誤差を，大阪市内において測定した結果を表 1 に引用する [6]．なお，PHS を利用した位置データは，高速道路高架のインターチェンジやビルなどの障害物により電界強度の影響を強く受けており，図 1 に，その測定誤差の偏りの一例を示す．すなわち，何らかの位置データ補正手法が必要である．その後，2000 年 5 月以後，GPS の精度向上したことや，FCC の E911 に準拠した携帯電話を用いた位置サービスが開始されていることに注意しておきたい．

【数値地図整備に関わる問題】

数値地図の道路線分は，図式分類コード，線分タグ，道路個別番号，多重連結構造などで表現される [13]．ただし，数値地図における道路データは，道路中心線を表現するものであり，道路線形に関するデータや，一方通行や交差点の右折禁止等の道路法規に関わるデータを含まない．加えて，自動車，徒歩，タクシー，バス，電車，自転車等の各種交通

表 1: 各デバイスの平均誤差 (大阪市中央区中低層ビル街，単位 m)

装置	平均誤差	中央値	標準偏差	標本数
PHS	188.7	187.2	41.7	71
GPS	24.4	20.9	10.5	29
DGPS	16.8	17.8	5.7	29

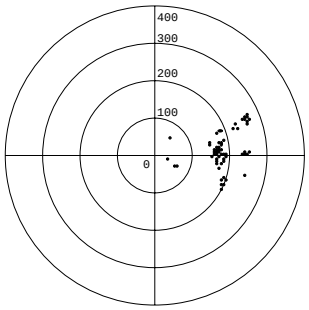


図 1: PHS の誤差分布 (大阪市中央区島町 1 丁目の交差点)[6] より引用

手段，さらに，時間帯に応じて，移動できる経路制約が変化するため，実用化をめざすための地図データには経路推定に適した諸属性の追加が求められる．

なお，複数組織のもつ複数の情報統合を推進することを目指す，地理データに対するクリアリングハウス構築も試みられている．そして，これらの実用化が進展すれば，我々の検討しているシステム構築コストが軽減されると期待している．しかしながら，各種の社会的制度やコスト負担の関係から判断するに，十分実用的であるデータ整備が実施されるまでにはかなり長い時間が必要であると予想している．

3　高次元インデックスと空間問合せ処理

本研究の目標とするシステム構築におけるデータベース技術として，移動体の位置データや地図データの蓄積を行う際に必要となる空間インデックスや時空間インデックスが必要であり，かつ，その利用目的に合致した時空間問合せを効率的に実行できることが不可欠である．そこで，本章では，空間データ，時空間データに対して提案されている各種高次元インデックスと問合せについて述べる．

3.1　空間・時空間インデックス

空間データを効率的に蓄積・検索する基本的技術として，データの空間的配置に基づいてバケット (bucket) と呼ぶ領域に再帰的に分割する階層的データ構造と空間インデックスがある．なお，蓄積の対象とするデータ型や分割規則などによって，多様な手法が提案されてきた [17, 18].

また，空間データそのものを，どのような目的で検索するかは，最も重要なポイントである．例えば，移動体から得られた位置データ周辺のオブジェクト検索は不可欠であるし，

また，デバイスの測定誤差を含むことを考えると，誤差半径内に存在するオブジェクト検索も必要である．すなわち，どのような種類の検索問合せを，頻繁に利用するかである．

そこで，この種の検索を効率良く実現するうえで，幾つかの空間インデックスとしてR-treeやQuadtreeなどを検討した．ただし，時々刻々と変化する多数の移動体の位置データ管理と解析を必要とする場合には，通常の空間インデックスでは，データ更新による負荷増大が問題となる．そのため，移動体の将来の位置を考慮し，時間軸に関する高次元化を行う時空間インデックスであるTPR-tree (Time Parameterized R-tree)[16]の適用を試みた[4]．また，3DR-Tree[19]の拡張について検討を行っている．

以下，幾つかの高次元インデックスを簡単に紹介する．

【Quadtree】

Quadtreeは，1つの位置データが，1つのバケットに対応するまで，バケットを再帰的に4等分する手続きを繰り返すことで木を生成する．従って，葉ノードには，位置データを含むものと含まないものが存在する．また，Quadtree構造の深さは最小で $\lceil \log_4 N \rceil$, 最大で $\lceil \log_2((s/d) \cdot \sqrt{2}) \rceil$ である．ここで，dはデータ間の距離の最小値である．すなわち，dが非常に小さい場合，木が深くなり検索効率が低下する．よって，1バケットに含む位置データ数の閾値 $c(> 1)$ を設定し，バケット内のデータ数を抑制しながら分割する手法も提案されている．

【TPR-tree[16]】

TPR-treeはR-tree構造に基づくが，各ノードに対応するd次元長方形の位置や大きさを時間関数で表現する． したがって，オブジェクトのグループ生成手法はR-treeと異なり，現在から将来にわたって位置が近いと推測できる位置データを同グループに索引付ける．

すなわち，TPR-treeは，時刻 $t = t_{ref}$ において，位置 $\mathbf{x}_{ref}$，速度 $\mathbf{v}$ をもつ移動体に対して，時刻 $t(t > t_{ref})$ における位置 $\mathbf{x}(t)$ を，(1) 式を用いて近似する．

$$\mathbf{x}(t) = \mathbf{x}_{ref} + \mathbf{v}(t - t_{ref}) \tag{1}$$

そして，位置$\mathbf{x}$，速度$\mathbf{v}$に対し，$(\mathbf{x}, \mathbf{v})$ 空間内の距離が近いものを同ノードに格納する．このことによって，境界長方形の時間経過による拡大を抑制した効率的な空間インデックスが生成できる．

【3DR-tree】

3DR-TreeもR-tree構造に基づく．ここで，時間軸方向の特性について注意深い深い検討が必要なことに注意しなければならない．図2に表すような時刻 t における定時刻問い合せや，短時間帯問合せを処理する場合に，ルートノードに近い親ノード操作が生じ，非効率なデータアクセスが生じる場合がある．また，長時間移動しないオブジェクトが存在する場合は，t 軸方向の長大ノードが生じ，その t 軸上の長大ノードを含む親ノードが生成されるため，やはり検索コストが高くなる．よって，対象データの時空間制約を踏ま

えたノード分割手法が鍵となる [19].

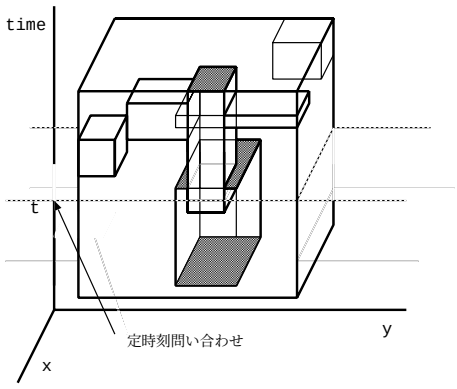


図 2: 3DR-tree の一例

ところで，時空間オブジェクトは，多くの属性をもつ．例えば，交通データを対象とするならば，道路の特性，走行車両の特性，人に付随する特性などであり，より高次元のデータを対象とした検索・解析処理が必要となる．もっとも，次元が高くなるにつれて，その処理が困難になると考えるのは妥当であろう．そのため，高次元データに対して，どのような検索・解析処理が必要となるのか，また，その計算コストがどう変化するかを深く検討するべきである．ただし，ここでは，今後検討を進める上で，文献 [10] で検討したクラスタリングアルゴリズムの研究，また，FastMap[5] 以後の高次元データに関する多くの研究が役立つと考えていることを述べるに留める．

また，先に述べたように，空間インデックスを利用することで，交通管制・情報提供においてより効果的な問合せ処理を実現できるか否かが重要な目標である．そこで，移動体から得られた実際の位置データに対する問合せ処理を試みている．具体性をもたせるため，対象となる位置データの実例として，移動体から実際に取得したデータ（表 2）と平面直角座標系に変換した位置データ（表 3）の一部を示す．これら携帯端末から収集されるデータは，ID，日時，携帯端末が探知したアンテナ数，位置座標 (緯度，経度)，速度の各座標成分，地図データベース内に建物情報がある建物内などの判定ビット，さらに，地点情報などで構成され，1 点あたり約 80 バイトとなる．

【位置検索処理と経路推定】

携帯端末から得られた位置データに測定誤差が存在することを考慮した上で，地図上の位置を決定しなければならない．そこで，位置データを中心とする誤差半径をもつ円を描く領域の検索処理が必要となる．もし，ここで，Quadtree を利用すると，誤差半径の円を被覆するバケット内のオブジェクト検索が実行できる．そして，それらのバケット内に格納された道路線分と道路ノードとの距離を求めることで，地図上の位置特定が合理的に行える．

表 2: 変換前の位置データ

経緯度座標 (単位：度)			
緯度	経度	ID	日時
34.680616	135.506702	1	1999/09/13 14:32:16
34.680661	135.506702	2	1999/09/13 14:32:33
34.681065	135.506425	3	1999/09/13 14:32:52
34.681939	135.506843	4	1999/09/13 14:33:10
34.681310	135.506096	5	1999/09/13 14:33:33

表 3: 変換後の位置データ

平面直角座標 (単位：10cm)			
x 座標	y 座標	ID	日時
7593	8031	1	1999/09/13 14:32:16
7643	8031	2	1999/09/13 14:32:33
8092	7780	3	1999/09/13 14:32:52
9060	8168	4	1999/09/13 14:33:10
8366	7480	5	1999/09/13 14:33:33

また，時間的に変化する連続した位置データが得られている場合，その移動経路推定が必要となる．例えば，携帯端末を中心とする円を用いた経路推定アルゴリズムである Screening 法が，朝倉らにより提案されている [20]．しかし，空間インデックスの性質を考慮すると，正方形 (誤差半径円に対する MBR (Minimum Bounding Rectangle)) を利用する方が効率的であろう．そこで，共通バケット内の道路線分と道路ノードへの距離が近いものを位置候補として検索し，移動経路推定を試みた．

【合理的移動経路推定】

より精度の高い経路推定を行うには，さまざまな移動体の特性にあわせながら，移動時間や滞在時間を考慮して，より合理的な移動経路を推定する必要がある．そのため，移動体から得られた位置データと，道路線分，道路ノードとの最短距離を求めて地図上で候補集合を生成し，さらに，これら候補集合に対して通過時刻を考慮しながら最短経路をダイクストラ法で求める．そして，交通手段や一方通行などの非空間属性による制約をも考慮した上で，合理的移動経路を推定する方法を提案した．このように，空間データ以外の属性値を用いることで，候補集合を抑制し，計算コストの軽減を試みることも必要である．

【交通工学と空間問合せ処理】

移動体に関わる各種空間問合せを，交通工学の視点から考える．なお，以下に述べる 3 種類の問合せにおいて，t_1，$t_2(t_1 < t_2)$ は時刻，R_1，R_2 は地図上の方形領域を表す．

1. **定時刻問合せ** $Q_{ts} = (R_1, t_1)$: 図 3 に示したように，ある時刻 t において，領域 R_1 内に存在する移動体を検索する．
   (例) 出力結果の集合は，交通密度や空間平均速度といった，ある瞬間の交通状況を表現する空間交通流パラメータ算定に必要となる．
2. **時間帯問合せ** $Q_{win} = (R_1, t_1, t_2)$: 図 4 に示したように，ある時間帯 t_1 から t_2 の間に領域 R_1 内に存在する移動体を検索する．
   (例) 空間平均速度に比較して安定した値をもつ時間平均速度などの時間交通流パラメータ算出に利用する．
3. **等速移動領域問合せ** $Q_{mov} = (R_1, R_2, t_1, t_2)$: 図 5 に示したように， 位置ベクトル $\mathbf{x}$ と時刻 t の 空間 $(\mathbf{x}, t)$ 内で，(R_1, t_1) と (R_2, t_2) とを結ぶ台形内に存在する移動体を検索する．
   (例) 進行方向と経過時間を考慮した時空間領域の交通密度を算出し，混雑状況の予測に利用する．

以上述べたように，多数の移動オブジェクト管理と，交通管制・情報提供に利用可能なシステム構成技術を検討してきた．基本となる空間データベースは，少なくとも，移動体から得られる位置データと道路地図データを格納しなければならない．そのため，道路データは空間インデックス (Quadtree) によって，移動体の位置データを時空間インデックス (TPR-tree) で索引付ける．加えて，蓄積した空間データに対して，各種時空間問合せ処理による解析を実行することを考えると，より高度なデータ構造が必要であると予想される．

4 今後の課題

本稿では，高次元インデックス技術の面から，携帯端末から得られる位置実測データを用いた合理的移動経路推定，交通工学の視点から交通管制・情報提供に必要な問合せ処理方式を実現する位置情報システム構築について考えた．このように，実システムにおける高次元データに対して，どのような検索・解析処理が必要となるのか，また，その計算コストがどのように変化するかの検討を進めることを重視している．また，空間属性と非空間属性を併せもつ大量の時系列高次元データに対して，交通計画やマーケティングの問題に適用可能なデータマイニング技術を発展させることは目標に含まれる．今後，より大きな母集団から長期間の経路推定データを収集し，現実問題へと対処できる高次元データ処理技術の検討を進める予定である．

謝辞

熱心な討論とデータ提供を頂いている田名部 淳 氏（都市交通計画研究所），実験データ処理のサポートを頂いた南 卓朗，岸 浩史 両氏（京都大学）に心より感謝の意を表す．また，南山大学 長谷川 利治 教授，愛媛大学 朝倉 康夫 教授の研究グループ，都市交通計画研究所，地域未来研究所の皆様にも深く感謝の意を表す．

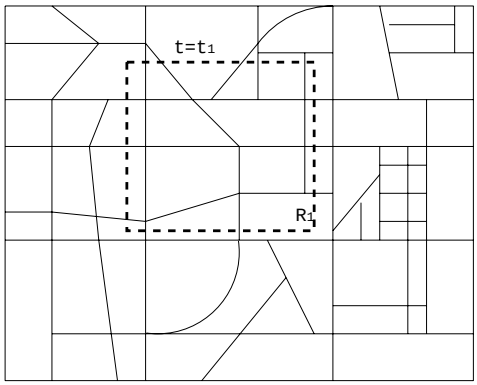


図 3: 定時刻問合せの例

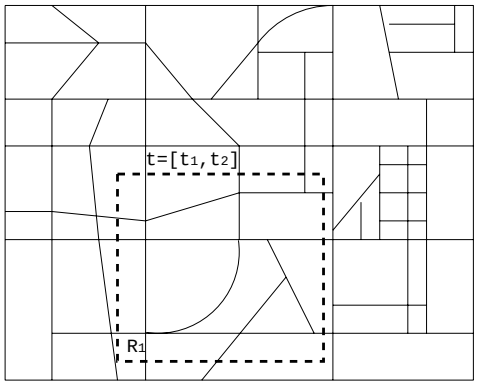


図 4: 時間帯問合せの例

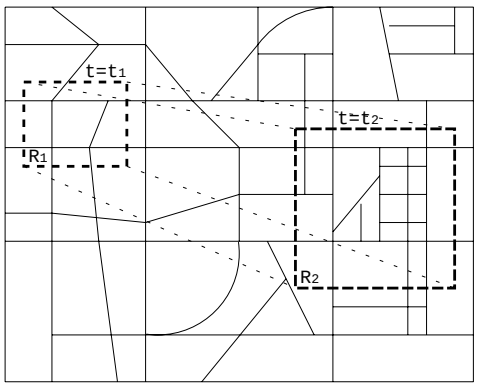


図 5: 等速移動領域問合せの例

参考文献

[1] 朝倉康夫, 羽藤英二, “時空間アクティビティデータ収集のための移動体通信システムの有効性に関する基礎的研究,” 交通工学, Vol. 35, No. 4, 2000.

[2] 朝倉 康夫, 羽藤 英二, 大藤 武彦, 田名部 淳, “PHS による位置情報を用いた交通行動調査手法”, 土木学会論文集, No.653/IV-48, pp.95–104, 7, 2000.

[3] DBS 研究会, “特集:空間メディアと GIS，および一般,” 情報処理学会研究報告, 2000-DBS-120, 2000.

[4] Ester, M., Frommelt, A., Kriegel, H.-P., and Sander, J., “Algorithms for Characterization and Trend Detection in Spatial Databases”, Proc. of the fourth ACM SIGKDD International Conference (KDD-98), pp.44-50, 1998.

[5] Faloutsos, C., and Lin, K., “FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets,” Proc. of 1995 ACM SIGMOD International Conference on Management of Data, pp.163-174, 1995.

[6] 阪神高速道路公団, “移動体位置検索技術の動向”, 移動体情報の活用に関する研究会資料, 2 月, 2000.

[7] (財) 阪神高速道路管理技術センター, “阪神高速道路交通管制施設に関する調査検討業務報告書,” 阪神高速道路交通管制施設に関する調査検討委員会, 2001.

[8] (財) 阪神高速道路管理技術センター, “画像センサに関する検討,” 阪神高速道路の交通管制に関する調査研究委員会, 2001.

[9] 伊藤 穣, 河野 浩之, 長谷川 利治, “空間データマイニングにおけるクラスタ発見とインデックス構造の利用,” 人工知能学会全国大会論文集, pp. 231–234, 1996.

[10] 川原 稔, 河野 浩之, 長谷川 利治, “空間データ発掘によるクラスタ発見手法の精度評価,” 情報処理学会第 53 回全国大会講演論文集 (3), pp. 19–20, 1996.

[11] 河野 浩之, “位置情報活用に関する基礎的考察〜経路データ活用とインデックス〜 ,” 情報処理学会研究報告, 2000-DBS-122, pp.291–298, 2000.

[12] Koperski, K. and Han, J., “Discovery of Spatial Association Rules in Geographic Information Databases,” Proc. 4th International Symposium SSD ’95, pp. 275–289, 1995.

[13] (財) 日本地図センター, (http://www.jmc.or.jp).

[14] 日本建築学会編, “建築・都市計画のためのモデル分析の手法,” pp.71–84, pp.122–162, 井上書院, 1992.

[15] Rogers, S., Langley, P., and Wilson, C., “Mining GPS Data to Augment Road Models,” Proc. of the fifth ACM SIGKDD International Conference (KDD-99), pp.104–113, 1999.

[16] Saltenis, S., Jensen, C. S., Leutenegger, S. T., and Mario A. Lopez, “Indexing the Positions of Continuously Moving Objects,” Proc. of the 2000 ACM SIGMOD International Conference on Management of Data, pp.331-342, USA, 2000.

[17] Samet, H., “Spatial Data structures,” Modern Database Systems, (W. Kim, ed.), ACM Press, New York, pp. 361–385, 1995.

[18] Samet, H., “The Design and Analysis of Spatial Data structures,” Addison-Wesley, Reading, Mass. New York, 1995.

[19] Tao, Y. and Papadias, D., “The MV3R-Tree: A Spatio-Temoral Access Method for Timestamp and Interval Queries,” *Proc. of VLDB 2001*, pp.431-440, Sep. 2001.

[20] 都市交通計画研究所, 地域未来研究所, “移動体位置情報に基づく動的時空間データ分析技術の開発（技術開発報告書）,” 平成 10 年度新規産業創造技術開発費補助事業, 2000.

発表論文

[1] 河野浩之, 位置データ問合せ処理のための空間インデックス手法の検討第 45 回システム制御情報学会研究発表講演会, 2001.

[2] Kawano, H., “Architecture of Trip Database Systems: Spatial Index and Route Estimation Algorithm,” XIV International Conf. of Systems Science, Vol. III, pp.110-117, Poland, 2001.

[3] 河野浩之“ 位置情報システムにおける空間データ利用に関する検討, ”人工知能学会 FAI-46&KBS-54 研究会, pp.159-164, 函館, 2001.

[4] 南 卓朗, 田名部 淳, 河野 浩之, “空間インデックスを用いた移動オブジェクト管理システムの構成と性能比較,” 情処研報 Vol.2001, No.70, DBS-125, pp.225–232, 夏のデータベースワークショップ (DBWS2001), 函館, 2001.

A02 班：ユーザ指向アクティブマイニング

Beamwise Graph-Based Induction による構造データからの知識発見

研究代表者　元田 浩　(大阪大学産業科学研究所)
研究分担者　鷲尾 隆　(大阪大学産業科学研究所)
研究分担者　吉田 哲也　(大阪大学産業科学研究所)
研究協力者　松田 喬　(大阪大学産業科学研究所)

目的

近年の計算機ハードウェア性能とネットワーク技術の急速な進歩により，電子的に可読なデータの量も増加の一途をたどっている．それに伴い，膨大な蓄積データから有用な知識を発見することを目的とするデータマイニングの新しい手法が鋭意研究開発されており，さまざまな分野で多大な成果をあげている．しかし，現在の研究の多くは巨大なデータを扱えるものの，通常のデータベースを念頭においたものであり，複雑な構造を有するデータからの知識発見に関する研究はあまり多くない．

構造を有しない通常のデータに対して，現在非常に広く行われている方法はデータを属性とその値のペアで表現し，属性の値と同定したいクラスの関係を決定木 [7, 8] や分類規則 [4, 5] で表現するものである．データマイニングでよく使われる相関規則 [1] もこの表現形式に入る．しかし，属性と値のペアの表現形式は複雑な構造データを表現するのには適しておらず，問題によってはより強力な表現形式が必要となる．一方，帰納論理プログラミング [6] は一階述語論理表現を用いているため一般的な関係を表現でき，かつデータ表現と帰納された知識表現がともにホーン節であるため，獲得した知識を背景知識に加え，そのまま使うことができるという利点がある．しかし，強力ではあるが，誰もが手軽に使えるほど技術的に成熟していない．一般に個別の知識や概念はグラフ構造で記述でき，その取り扱いは非常に直感的で容易である．本研究で扱うグラフ表現は，表現能力の点では一階述語論理に劣るが，グラフ構造データからの知識発見は応用範囲も広く極めて重要な研究課題である．本特定領域研究で用いる共通医療データも多くの属性の時系列変化を扱ったものであり，属性間の時間的な因果関係はグラフ構造データとして表現できる．

グラフ構造データから特徴的なパターンを抽出することを目的としたアルゴリズムとして著者らは Graph-Based Induction (GBI 法) [2, 3] を提案してきた．GBI 法はグラフ構造データ中に現れる特徴的なパターンを，隣接する二つのノードを逐次的にチャンクすることによって発見することを目的として考案された Greedy 探索手法である．対象とするグラフは，ノード，リンクにラベルがあり，多入力・多出力，ノード間にループ（自己ループを含む）を許す一般有向・無向グラフである．本研究では，クラス分類能力のあるパターンを「特徴的なパターン」と定義した上で，より多くの「特徴的なパターン」をグラフから抽出できるように GBI 法に Beam 探索法を取り入れ，Greedy 探索法の欠点を補った．UCI Machine Learning Repository [9] のデータに適用し，Beam 探索の評価・検討を行った結果を報告する．

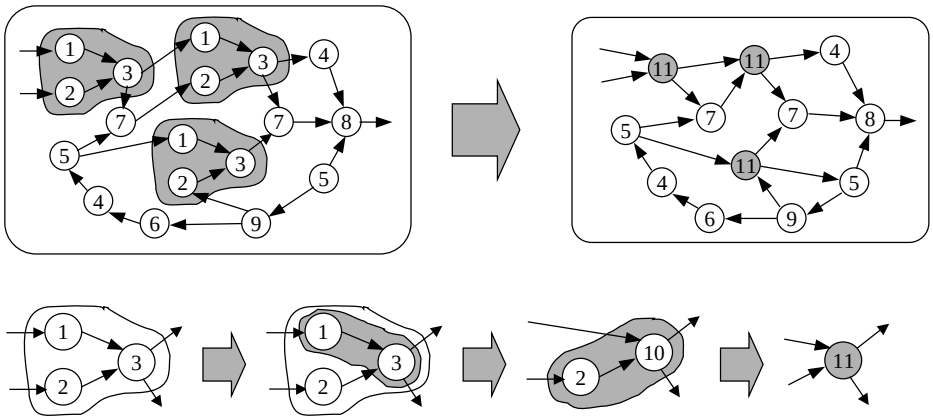


図 1: 逐次ペア拡張の基本アイデア

Graph-Based Induction

GBI 法の概要

GBI 法は，グラフ構造データ中に現れる特徴的なパターンを抽出することを目的として考案された [2]．*GBI* 法は 図 1 に示すように「ペアの逐次抽出（チャンク）により特徴的なパターンを抽出する」という基本アイデアにより実現されている．ここで,「ペア」とは「二つのノードおよびそれらをつなぐリンク」からなる *GBI* 法で用いる基本単位となるものである．また,「ペア」は「逐次拡張」されることにより複雑なパターンを構成するので，逐次拡張によりデータから抽出されたものを「抽出パターン」と呼ぶ．*GBI* 法が行う「逐次ペア拡張」は次の 3 ステップを繰り返し行うことで実現される．

Step 1. グラフ中に存在するペアをすべて抽出する．

Step 2. 抽出したペアのうち，ある評価指標によりチャンクすべきペアを一つ選び「抽出パターン」として登録する．このとき，ペアを構成するノードがすでに書き換えられたノードであれば元のパターンに復元してから登録する．チャンクすべきパターンがなくなれば終了する．

Step 3. **Step 2**. で選ばれたペアを一つのノードの置き換えることにより，グラフを書き換える．**Step 1**. に戻る．

入力グラフに対して上記 3 ステップを繰り返すことで，ペアを逐次拡張し，最終的に登録された「抽出パターン」の集合としてデータに含まれる「特徴的なパターン」の集合を得ることができる．ただし，**Step 1**. でペアを抽出する際，一つのノードに書き換えられたパターンを元のパターンに復元しての評価は行わない．つまり，探索空間を制限するという点で，*GBI* 法は探索手法としては Greedy 探索手法である．そのため，入力データ中に存在するすべての「特徴的なパターン」を抽出することはできない．なお,「特徴的なパターン」は対象とするドメインごとに様々に定義することができるが，従来は頻度を用いていた．本研究では「特徴的なパターン」をクラス分類能力を有するパターンと定義する．頻度と違いクラス分類能力を評価指標は単調性が保証されないため，頻度以外の

指標をペアの選択に用いると後述する理由で所望の結果が得られない可能性がある．この問題を回避するために，*GBI* 法のアルゴリズムを改良した．

アルゴリズムの改良

GBI 法では，それぞれの繰り返しにおいて書き換えられたグラフから得ている情報は，そのグラフに存在するペアのうちどのペアが最も評価指標の値が大きくなるかということのみである．得られる「特徴的なパターン」は逐次拡張されたパターンなので，そのパターンにたどり着くまでの部分構造がすべて各時点で数えられたペアの内で評価指標が最大（あるいは最小）であるという条件を満足している必要がある．つまり，頻度などの単調な評価指標ではうまく機能するが，非単調な場合はうまく機能しない．そこで，チャンクすべきペアを決める指標としては単調性を有する頻度 (単調性を有するものであれば，必ずしも頻度である必要はない）を用い，それとは別に，グラフから「特徴的なパターン」を抽出するための評価指標を用いることができるように *GBI* 法を改良した．後者の指標は非単調性を有していても構わない．改良後の *GBI* 法は以下の 4 ステップを繰り返すことでグラフから「特徴的なパターン」を抽出する．

Step. 1 グラフ中に存在するペアをすべて抽出する．

Step. 2a **Step. 1** で抽出したペアのうち，評価指標により「特徴的なペア」をすべて登録する．このとき，ペアを構成するノードがすでに書き換えられたノードであれば元のパターンに復元してから登録する．

Step. 2b **Step. 1** で抽出したペアのうち，頻度指標によりチャンクすべきペアを一つ選び,「抽出パターン」として登録する．このとき，ペアを構成するノードがすでに書き換えられたノードであれば元のパターンに復元してから登録する．チャンクすべきペアがなくなれば終了する．

Step. 3 **Step. 2b** で選ばれたペアを一つのノードに置き換えることにより，グラフを書き換える．**Step 1.** に戻る．

この改良された *GBI* 法の出力は **Step. 2a** で抽出された「特徴的なパターン」の集合である．この改良により，うまく評価指標を選択することであるクラスにのみよく現れるパターン，あるいはあるクラスにのみ現れないパターンなどを抽出することができるようになると考えられる．

Beamwise *GBI* 法

GBI 法はすでに述べたようにペアの逐次拡張という Greedy 探索手法を採用している．グラフからすべての部分グラフを抽出する問題は NP 完全問題であるため，すべての「特徴的パターン」を抽出するのではなく，ある程度の大きさを持った有意な部分グラフを抽出することを目的としている．そのため，大規模なグラフ構造データから「特徴的なパターン」を抽出するには非常に有用である．しかし， Greedy 探索にも問題がある．すべてのノードのラベルが異なる場合はチャンクするペアにあいまい性は生じないが，同

じノードラベルが多数存在している場合には，評価値が同値のペアや同種類のペアの連鎖[1]が生じるためにチャンクすべきペアの選択にあいまい性が生じる．

そこで，*GBI* 法に Beam 探索を導入することで，この問題の低減を図る．具体的には，チャンクするペアを唯一ではなくある一定の数だけ選択し，それぞれのペアについてチャンクする．これにより，複数の並列な状態に分裂するが，次のステップではそれぞれの状態について，ある一定の数だけのチャンクすべきペアを選択するのではなく，すべての状態の中で上位のある一定の数のパターンを選択しチャンクすべき候補とする．これにより，チャンクのステップが進んでいくにつれて，状態が爆発的に増加するのを防ぐことができる．以上すべてのことを考慮した Beamwise *GBI* 法のアルゴリズムは以下のとおりである．

Step. 1 すべての状態について，グラフ中に存在するペアをすべて抽出する．

Step. 2a **Step. 1** で抽出したペアのうち，評価指標により「特徴的なペア」をすべて登録する．このとき，ペアを構成するノードがすでに書き換えられたノードであれば元のパターンに復元してから登録する．

Step. 2b **Step. 1** で抽出したペアのうち，頻度指標によりチャンクすべきペアをある一定の数だけ選び,「抽出パターン」として登録する．このとき，ペアを構成するノードがすでに書き換えられたノードであれば元のパターンに復元してから登録する．チャンクすべきペアがなくなれば終了する．

Step. 3 **Step. 2b** で選ばれたそれぞれのペアに対し，ペアを一つのノードに置き換えることにより，それぞれのグラフを書き換える．この際，必要に応じて状態を分裂・消去させる．**Step 1.** に戻る．

Beam 幅を 5 に設定した場合の状態遷移例を 図 2 に示す．まず初期状態は 1 つの状態 cs から始まる．状態 cs 中に含まれるペアをすべて数え上げ，評価指標により「特徴的なパターン」を抽出し，チャンクすべきペアを頻度指標が上位のものから 5 つ「抽出パターン」として選択する．それぞれの抽出パターンに基づき，グラフをチャンクし書き換える．このとき，抽出パターンは 5 つ選択されているために初期状態 cs は 5 つの状態（c11 〜 c15）に分裂させる．次に，すべての状態についてグラフ中に存在するペアを数え上げ，評価指標により「特徴的なパターン」を抽出し，さらに頻度指標によりチャンクすべきペアを 5 つ選択する．状態 c11 からはチャンクすべきペアが 2 つ選ばれているためにグラフを二つに分裂させ，チャンクすべきペアそれぞれに基づいてそれぞれグラフを書き換える．また，状態 c12 ではペアが一つも選択されなかったため，消滅する．このことを，終了条件を満足するまで繰り返し行う．

このように探索空間を増加させることにより，*GBI* 法が持つ特徴的なパターンを抽出する能力が向上する．

[1] $a \to a \to a$ といった構造の場合，どちらの $a \to a$ をチャンクすべきかを決定できない．

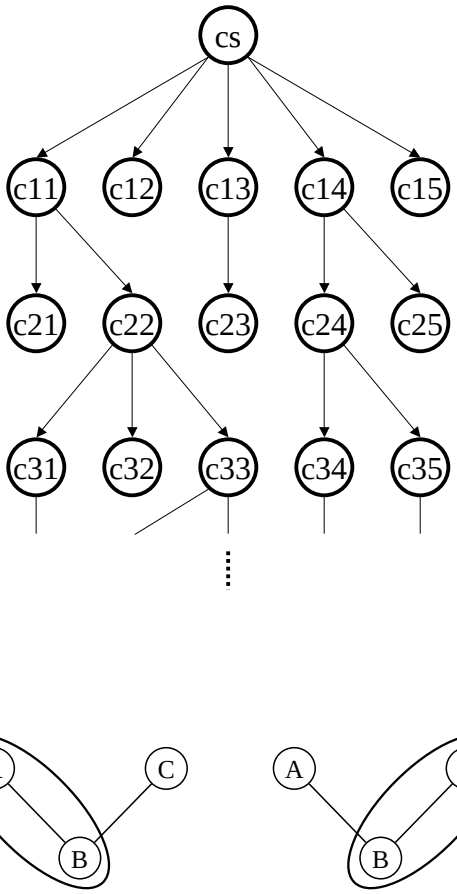


図 3: 同じグラフパターンを表すことなるペア

Canonical Label

GBI 法は，パターンをチャンクして一つのノードに書き換えるために，ペアを数え上げるときに同じグラフ構造を表しているにもかかわらず，異なったペアとして扱ってしまう場合がある．例えば，図 3 は同じグラフ構造を持っているが，チャンクされているために，ペアとしては別のものとして扱われる．

これを防ぐためには，ペアを数え上げるときにペアが表すグラフ構造が同じグラフを表すものであるあるかをチェックする必要がある．そのため，グラフを Canonical Label [10, 11] で表現し，Labelが同じものを同じグラフ構造と判定する．Canonical Label の基本的な作成方法は以下のとおりである．まず，グラフに含まれる頂点をそのラベルと次数（頂点につながる辺の数）でグループ分けし，辞書順に並べる．この辞書順に並べられた頂点に従って，隣接行列を作り，無向グラフの場合は上三角行列を縦（あるいは横）に連結してコード化し，これをグラフのコードとする．隣接行列とは，グラフ G 中の頂点 $v_i \in V(G)$ と頂点 $v_j \in V(G)$ が連結されている場合にその第 i 行第 j 列の要素が 1 となる行列のことである．ラベルと次数が同じ頂点が複数存在する場合は，それらの順列のうち，最も大きい（あるいは最も小さい）コードを Canonical Label として採用する．Canonical Label を用いることで，M 個の頂点を持つグラフがあり，その頂点が N

グループに分かれたとすると，それぞれのグループ内の頂点の数を $p_i(i = 1, 2, \ldots, N)$ として，$M!$ の探索空間を $\prod_{i=1}^{N}(p_i!)$ に減少させることができる．上三角行列を縦に連結した場合のグラフの隣接行列のコードは，以下のように定義される．

$$
\begin{aligned}
A &= \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ & a_{22} & \ldots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} \\
code(A) &= a_{11}a_{12}a_{22}a_{13}a_{23}\ldots a_{nn} \qquad (1) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{j}(2^{\sum_{k=j+1}^{m} k+j-i}a_ij) \qquad (2)
\end{aligned}
$$

この探索空間を次の手法によりさらに効率よく枝刈りする．まず，コードを求める際，隣接行列を縦に連結するか横に連結するかの 2 通りが考えられるが，縦に連結する手法を採用する．縦に連結すれば，上位レベルの頂点に関連する隣接行列の要素は常にビット列の上位にくるので上位のレベルの頂点の隣接行列での位置が決まれば，コードの上位ビットも同時に決定され，下位のレベルの頂点の順列には影響されないためである．たとえば，式 (1) において，順位 1 番目のノードと 2 番目のノードのみで決定できる要素は $code(A)$ の上から 3 番目のビットまでであり，この中にこれらより下のレベルのノードで決定できる要素は含まれていない．これにより，$\prod_{i=1}^{N}(p_i!)$ の探索空間を $\sum_{i=1}^{N}(p_i!)$ に減少させることができる．

しかし，まだこの手法では化学構造式のように頂点に同じラベルが多数存在し，頂点のラベルが決まればその頂点から出る辺の数もほぼ決定するような場合には，$\sum_{i=1}^{N}(p_i!)$ における p_i の値が大きくなるため，組み合わせ爆発を起こす．そこで，すでに決定している上位の頂点の情報を用いることでさらに探索空間を減少させる手法を述べる．あるグラフ G において，すでに上位の頂点 $v_i \in V(G)(i = 1, 2, \ldots, N)$ が決定していたとする．このときラベルと次数が同じである頂点集合 $u_i \in V(G)(i = 1, 2, \ldots, k)$ において，これらの頂点の順列のうち最もコードが大きくなる頂点の順列を求めることを考える．v_{N+1} になるべき頂点は頂点 u_i の内頂点 v_1 と連結している頂点である．なぜなら，頂点 v_{N+1} が関係するコードのビットの内，最も上位にあるビットは a_{1N+1} なので，このビットを 1 にする頂点，つまり v_1 と連結する頂点が最もコードを大きくする頂点となるためである．頂点 u_i の内，v_1 に連結する頂点が複数あったり，あるいは連結する頂点がまったくない場合はそれら（まったくない場合はすべての頂点）の内，先ほどと同様の理由で，v_2 と連結している頂点が v_{N+1} になるべき頂点である．このことを繰り返すことで，v_{N+1} になる頂点を決定することができる．v_N まで比較しても v_{N+1} の頂点が決定できない場合はこれらは候補として残す．

例えば，図 4 において，すでに 1, 2, 3 番目のノードは決定しているとする．このとき，4 番目のノードとなるのは 1 番目のノードとつながるノードラベル d のノードと決定できる．なぜなら，すでにコードの上位にくるビットは決定しており，この状態でコードを最大にするためには最も上位のノードである 1 番目のノードとつながるノードを上

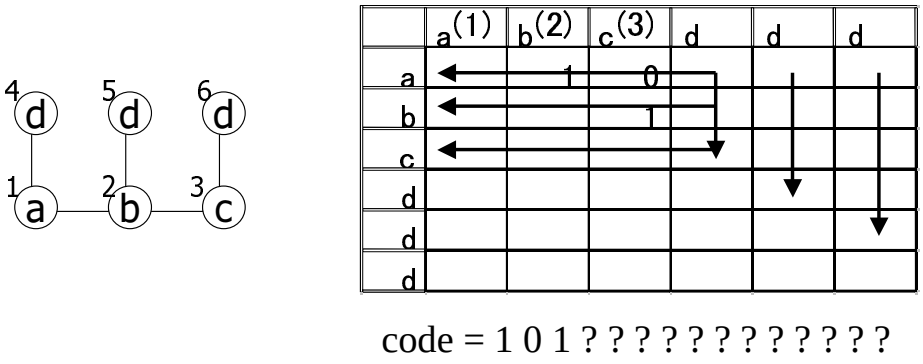


図 4: 同じレベルのノードの決定

位にもってくるべきだからである．同様に， 5 番目のノードは 2 番目のノードとつながるノードラベル d のノード， 6 番目のノードは 3 番目のノードとつながるノードラベル d のノードと決定することができる．この結果，ノードの順列をとる必要が無く探索空間を減少させることができる．これにより，$u_i(i = 1, 2, \ldots, k)$ の順列 $k!$ の探索空間をすべて探索する必要がなくなり，たとえば l 個の頂点がこの方法で決定できた場合には， $k!$ の探索空間を $(k - l)!$ に減少させることができる．

実験と考察

前節で示したアルゴリズムを実装し，評価実験として DNA の塩基列データに適用した．ドメインは DNA の塩基列データより，クラス分類に適したパターンを抽出することである．用いたデータセットは UCI Machine Learning Repository [9] により提供されている Promoter データセットである．Promoter データセットは塩基をあらわす A，G，T，C からなる長さ 57 の文字列データであり，クラスはその塩基列が “Promoter” [2]を含むことを示す Positive と “Promoter” を含まないことを示す Negative の 2 つである．データセット中の事例数は 106 個で，クラス Positive のデータが 53 個，クラス Negative のデータが 53 個である．

これを通常の属性とその値のペアとしてデータ表現すれば，簡単に分類器を構築できるが，そのようなデータ表現では，配列順に属性を番号づけなければならず，塩基の相対配列の特徴を取り出すことができない．従って，グラフ表現が有効となる．本研究ではこの文字列を 図 5 のように塩基をノードラベルとし，一つの塩基から 1 〜 10 個先の塩基にそれぞれ 1 から 10 のラベルをつけたリンクを張ることで一つのグラフに変換し， GBI 法への入力とした．１つのグラフのサイズはノード数 57 個，辺の総数 515 本となり，かなり大きなグラフでとなる．

チャンクするためのペアを選択する頻度指標の閾値は 20%とした．すなわち，ある状態において 20 %以上 (つまり 22 個以上) のグラフに含まれているペアで頻度の多いものからチャンクすべきペアを選択した.「特徴的なペア」を抽出するための評価指標には正規

[2]DNA を鋳型に mRNA 合成を開始する DNA 上の特定の塩基列を指す.

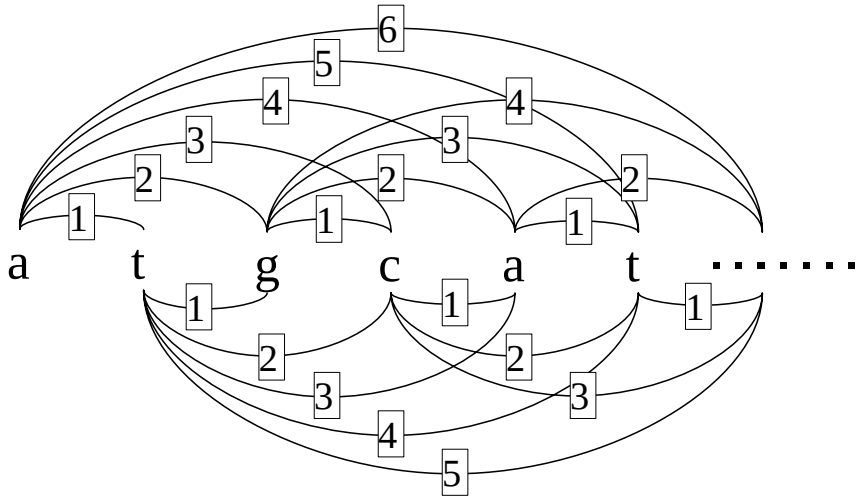


図 5: DNA データからグラフへの変換

化した確率を表す

$$Max.\left\{\frac{\frac{p}{P}}{\frac{p}{P}+\frac{n}{N}},\frac{\frac{n}{N}}{\frac{p}{P}+\frac{n}{N}}\right\} \tag{3}$$

を用い，この値が閾値以上になるものを「特徴的なパターン」と定義した．ここで，p および n はそれぞれ評価対象のパターンを持つクラスが Positive あるいは Negative のグラフの数，P および N はクラス Positive あるいは Negative のグラフの数（それぞれ 53）である．式 (3) の閾値には 0.6, 0.7, 0.8 の 3 通りを使用した．Beam 幅は 1 と 5 〜 50 まで 5 刻みに設定した．得られた「特徴的なパターン」の数とパターンに含まれる平均ノード数を 表 1 に示す．次に，得られたパターンの評価を行うために，得られたパターンを用いて分類規則学習を行った．各パターンのあり，なしの 2 値をとる属性とみなし，これらを用いてそれぞれのデータを表現し，これを入力データとして学習した．

分類規則学習アルゴリズムでは標準的な手法である C4.5 [8] を用い，leaving-one out 法によりエラー率を測定した．結果を 図 6 に示す．この結果より，Beam 幅 25 程度まではエラー率が減少傾向にあることが分かる．最大約 10 % 程度まで下がっていることからも，クラス分類に寄与する属性（パターン）が発見されていることが分かる．その後，Beam 幅の増加とともにエラー率も増加していくが，これは Beam 幅が増えることで，探索空間のより深い所まで探索され，このデータセットに特化した一般的ではないパターンが抽出されやすくなったためだと思われる．つまり，Beam 幅 25 付近を境目に過学習状態に入ったと考えられる．ちなみに，先に述べた属性と値のペアでデータ表現し同じ C4.5 で構築した決定木の leaving-one out 法によるエラー率は 16.0 % であり，本手法が有効に機能していることが例証される．

今回の実験では，過学習まで考慮に入れなかったが，これを防ぐためには，あまりデータセットに特化したパターンは抽出しないよう評価指標を工夫したり，一旦属性を生成した後，フィルター法で属性選択をする必要があると思われる．しかし，今回の実験により，構造を有するデータを直接，属性・属性値のペアでデータ表現したのでは求められない規則表現が，GBI 法により属性構築することで，同じ属性・属性値のペアのデータ表現を

Beam 幅		1	5	10	15	20	25
閾値	パターン数	355	1442	2125	3174	3688	4733
0.6	平均サイズ	3.8	3.9	3.8	3.8	3.9	3.9
閾値	パターン数	81	282	399	590	645	896
0.7	平均サイズ	3.9	4.1	4.0	4.0	4.1	4.0
閾値	パターン数	16	41	38	72	73	117
0.8	平均サイズ	4.1	4.3	4.0	4.1	4.1	4.2

Beam 幅		30	35	40	45	50
閾値	パターン数	4697	5604	6293	6780	7342
0.6	平均サイズ	3.8	3.9	3.9	3.9	3.9
閾値	パターン数	815	992	1136	1162	1249
0.7	平均サイズ	4.0	4.1	4.0	4.1	4.1
閾値	パターン数	115	145	163	161	178
0.8	平均サイズ	4.2	4.3	4.2	4.2	4.3

表 1: 得られたパターンの数

念頭に置いて設計された通常の分類規則学習法でも求められることが示された.

まとめ

GBI 法のアルゴリズムを改良し Bean 探索を導入することで，構造データから「特徴的なパターン」を抽出する能力の向上を図ることができた．また，*GBI* 法を属性構築に用いることで，一般によく用いられる属性・属性値のペアを念頭に置いた分類規則学習法でも構造を有するデータから分類規則学習法を行えることを示した．今後は，評価指標等を改良することで，過学習状態に陥らず，さらによい「特徴的なパターン」を抽出する能力を高めていくと共に，本手法を共通医療データからのマイングに適用して行きたい.

参考文献

[1] R. Agrwal and R. Srikant. First Algorithms for Mining Association Rules. Proc. of the 20th VLDB Cnference, pp. 487–499, 1994.

[2] 吉田 健一, 元田 浩. 逐次ペア拡張に基づく帰納推論. 人工知能学会誌, Vol. 12, No. 1, pp. 58–67, 1997.

[3] 松田 喬, 元田 浩, 鷲尾 隆. 一般グラフ構造データに対する Graph-Based Induction とその応用. 人工知能学会誌, Vol. 16, No. 4, pp. 363–374, 2001.

[4] R. S. Michalski. Learning Flexible Concepts: Fundamental Ideas and a Method Based on Two-Tiered Representaion. In Machine Learning, An Artificial Intelligence Approiach, Vol. 3, pp. 63–102, 1990.

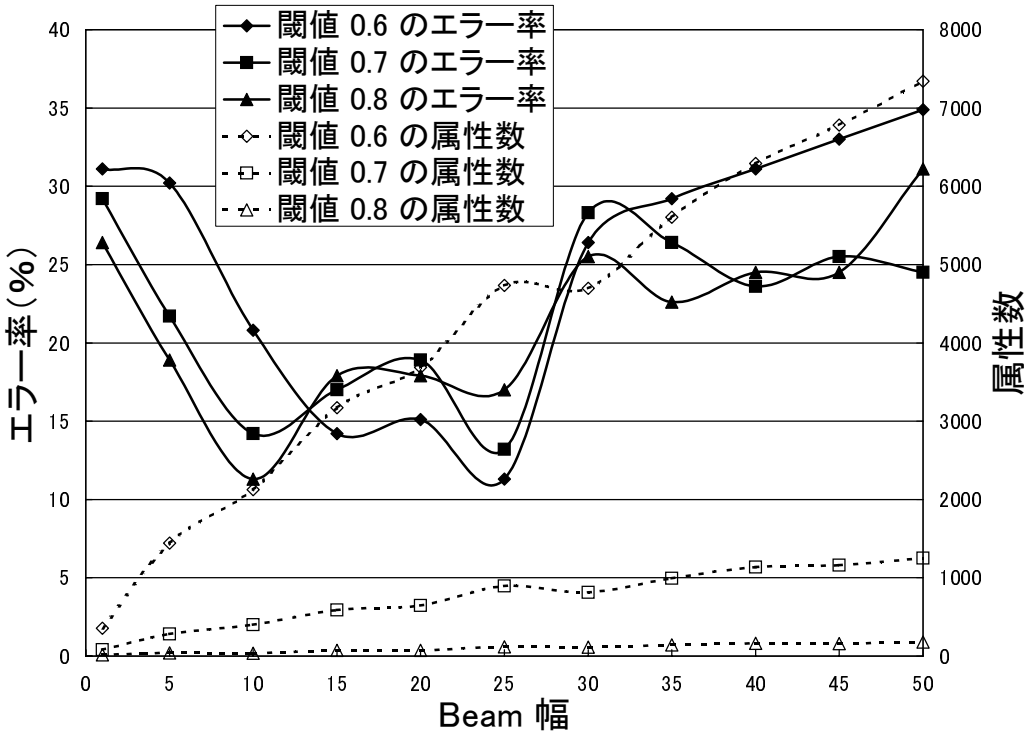


図 6: 分類規則学習の結果

[5] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone. The CN2 Induction Algorithm. Machine Learning, Vol. 3, pp. 261–283, 1989.

[6] S. Muggleton, S. and L. de Raedt. Inductive Logic Programming: Theory and Methods. Journal of Logic Programming, Vol. 19, No. 20, pp.629–679, 1994.

[7] J. R. Quinlan. Induction of decision trees. Machine Learning, Vol. 1, pp. 81–106, 1986.

[8] J. R.Quinlan. C4.5:Programs For Machine Learning. Morgan Kaufmann Publishers, 1993.

[9] Blake, C. L. and Keogh, E. and Merz, C.J., UCI Repository of Machine Leaning Database, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[10] R. C. Read and D. G. Corneil. The graph isomorph disease. Journal of Graph Theory, 1, pp. 339–363, 1977.

[11] S. Fortin. The graph isomorphism problem. Technical Report 96-20, University of Alberta, Edomonton, Alberta. 1996.

ブランド間関連購買の知識表現と評価基準

研究分担者　矢田勝俊（関西大学商学部）
研究協力者　羽室行信（大阪産業大学経営学部）
加藤直樹（京都大学大学院工学研究科）

背景と目的

現代のコンピュータ技術の急激な発展によって、日常業務の自動化が大きく進み、データベースに膨大なビジネスデータが蓄積されるようになった。しかしながら、多くの企業は将来の自社の戦略計画にそのような蓄積されたデータを十分に活用していないようである [13, 14]。近年、日本の流通業において顧客の膨大な購買履歴が蓄積されつつある。もし、これらの大量のデータから有用な知識が発見されれば、より有効なマーケティング戦略が策定可能になる [2, 3, 4, 5, 7]と考えられている。

一方で、データベースからの知識発見、もしくはデータマイニングがビジネスデータから自動的に意味ある知識を抽出する新しい技術や方法論を扱うものとして、注目される研究領域になっている [1, 10, 11]。

そうした顧客に関するデータの中で、日本の流通業界で最も注目を集めているのが、ID 付 POS データと呼ばれる顧客の ID が認識できる POS 販売明細データである。従来、顧客の消費意識などに関する情報は、各企業や研究機関が行うアンケートを用いて収集するか、少数のパネル顧客と契約し、かれらの消費データを詳細に蓄積することなどで集められてきた。同時にそれらは、利用する多くのマーケターにとって十分な情報ではないことも認識されていた。ところが近年、FSP に代表される会員カードシステムが多くの流通業で導入され、顧客 ID の認識が可能な販売明細 POS が多様な業種、企業において蓄積されるようになった。実際にこれらのシステムは、データウェアハウスなどの莫大なシステム維持費がかかり、そのコストを短期に回収できるものではない。長期の顧客との関係を維持するための 1 つの手段である。しかし経済不況が深刻化する中で、これらのデータを積極的にビジネスに利用し、既存のビジネスを効率化しようとする動きが非常に多く見られるようになってきている。

そうした蓄積された POS データを分析するためのデータマイニングの基本的なテクニックの 1 つにアソシエーションルールがある。アソシエーションルールは、POS 端末で蓄積されたデータを分析することによって、関連購買のルールを発見するための典型的なものである。スーパーやドラッグストアにおいて、顧客は 1 度にいくつかのアイテムを購入する。POS 端末から得られるレシートを調査することによって、我々は、2 つの商品が同時にどれくらい頻繁に購入されるか、つまり同一レシート内に同時に最も頻繁に出現する商品の組み合わせを知ることができる。例えばこの購買関連性に関する情報は店舗内の商品の効果的なレイアウトに役立てることができる [2, 9]。

一般的に、そのような関連購買は 1 時点の購入機会だけではなく、長期にわたる複数購

入機会においても起こりうる。人間は購入に際しリスクを減らすために、以前、購入し評価した商品を継続購入する傾向がある。その最も典型的なパターンが特定ブランドの継続購入である。大きな付加価値をもたらし、さらに継続的な利益が維持できるこうした顧客を多くの企業が望んでおり、長期にわたる顧客行動の把握が重要なかぎになっている。

また、消費者の購買行為は通常、複数ブランド間の選択という状況で行われる。多くの小売店の店頭では無数の商品が並べられ、消費者はその中から様々な基準をもとに購入の意思決定を行っている。そうした状況のうち、ブランドシェア、つまりそのカテゴリにおけるあるブランドの占める割合は、重要な意味を持つ。当然、トップシェアのブランドは、それらのカテゴリを十分に習熟していない消費者に対して、購入される確率が高くなる。なぜなら、それらのブランドの店頭での販売面積は必ず他ブランドより広いからである。こうした状況を加味しながら、それでも特定ブランドに執着する顧客は存在する。

しかしながら、我々の知る範囲では、従来のアソシエーションルールが対象とする関連購買の分析を複数購入機会、複数ブランド間の関連性にまで拡張したものはない。我々はこうした拡張として新しい考え方、association strength を提案する。この概念によって、継続的な購入機会における複数ブランド間の関連性を表現することができ、新しい関連購買の評価基準を提示することができる。

マーケティングの分野において、アソシエーションルールから算出されるバスケット分析の結果は、これまで多くの知見を専門家に提供してきた。しかし、持続的な競争優位の確立を目指す企業にとって、長期間におけるライバルブランドとの関係の理解がますます重要になっており、本稿で焦点を当てた、複数購入機会を対象にしたブランド間の購買関連性を分析できる枠組みは重要な知見を提供できると考えられる。こうした特定の現実問題に対する具体的な知識表現、評価基準を確立は、アクティブマイニングの実現に不可欠である。

検討内容

ここでは、association strength の定義を紙おむつのケースを元にして説明する。サイズ M と L の両方の紙オムツを購入する顧客の集合を$C=\{1,2,...,n\}$、紙おむつのブランドの集合を$B=\{1,2,...,K\}$とする。顧客 i によって購入されたサイズ p のブランド j のオムツの枚数を$a_i^j(j)$とする。ただし、サイズ p は M か L とする。顧客 i が購入するサイズ p の紙おむつの購入枚数を$a_i^p \equiv \sum_{j=1}^{K} a_i^p(j)$とする。ここでブランドスウィッチが確率的に起こった場合、顧客 i の M サイズのブランド j から L サイズのブランド k へのブランドスウィッチ確率（$P_i(j,k)$）を考えてみる。顧客 i の L サイズのブランドシェアがそれぞれ等しいと仮定すると、$P_i(j,k)$は j に依存せず、

$$P_i(j,k)=a_i^L(k)/a_i^L \quad (1)$$

となる。つまり、ブランドシェアと等しい確率でブランドスウィッチが起こると推測され

る。これは購入頻度によって重み付けされた$P_i(j,k)$の 2 乗誤差を最小化する値 x として定義することができ、例えば x は以下の問題を解くことで計算される。

$$\min \sum_{i=1}^{n} a_i^M(j)(P_i(j,k)-x)^2 \quad (2)$$

ここでは、サイズMのブランドＪを大量に購入している顧客の購買行為が x の評価に強く反映しており、現実の市場をうまく表現していると考えられる。このようにして計算された x は顧客全体に対する M サイズのブランド j から L サイズのブランド k へのブランドスイッチ確率を表している。この値の L サイズにおけるブランド k のシェアとの相対的な大きさが M サイズのブランド j と L サイズのブランド k との相関の高さを表している。その値を次にブランドシェアと比較して association strength として定義する。

$$as(j,k)=P(j,k)/share^L(k) \quad (3)$$

$share^L(k)$は、L サイズの k のブランドシェアを表す。もし$as(j,k)$が 1 より大きければ大きいほど、M サイズのブランド j と L サイズのブランド k が高い関連性をもっていることを示している。

	サイズM			
	1	2	3	Total
Customer1	10	2	3	15
Customer2	15	4	1	20
Customer3	5	14	11	30
total	30	20	15	65
	サイズＬ			
	1	2	3	Total
Customer1	20	2	2	24
Customer2	10	3	3	16
Customer3	2	20	15	37
total	32	25	20	77

表 1 association strength の説明例

association strength の考え方をよりよく理解するために、表 1 のような 3 人の顧客、3 つのブランドで構成される実際の例で考えてみよう。表から顧客１と顧客 2 がサイズM、Ｌともにブランド１にロイヤリティを持っていることが見て取れる。したがって、association strength、$as(1,1)$は、高そうである。実際に計算してみると、$a_1^M(1)=10$、$a_2^M(1)=15$、$a_3^M(1)=5$、そして$P_1(1,1)=20/24$、$P_2(1,1)=10/16$、$P_3(1,1)=2/37$であるため、$P(1,1)=0.6$になる。こうして、$share^L(k)=32/77=0.416$から、$as(1,1)=0.6/0.416=1.44$が得られる。

上述の定義は、異なる商品カテゴリに属する 2 つのブランドの関係に一般化することが可能である。オムツの例を考えてみると、赤ちゃん用オムツの例では、2 つのブランドの購入には時系列的に順番が存在し、L サイズのオムツは必ず M サイズのオムツの後に買われるものである。しかしながら、一人の顧客は、長い購買履歴の中には、あるブランドと他のブランドとをともに購入しており、この場合、ブランド間に時系列の順番はない。この場合の association strength の定義もオムツのケースと同様な手順で求められる。

2 つの商品カテゴリ、C_1 と C_2 を考えてみよう。$B_1 = \{b_1^1, b_2^1, ..., b_{K1}^1\}$、$B_2 = \{b_1^2, b_2^2, ..., b_{K1}^2\}$ とする。そして C_1 のカテゴリのブランドを少なくとも 1 つ以上購入しかつ C_2 びカテゴリのブランドを少なくとも 1 つ購入した顧客の集合を $C = \{1,2,...,n\}$ とする。また、カテゴリ m=1,2 であり、顧客 i によって購入される C_m に属するブランド b_j^m の合計購入数量を $a_i^2(j)$ とする。顧客 i によって購入されるカテゴリ C_m に属するすべてのブランドの購入量、つまりそれぞれのカテゴリ全体の顧客ごとの購入量を $a_i^m \equiv \sum_{j=1}^{K_m} a_i^m(j)$ とする。ここでブランド選択が確率的に起こるものと仮定する。ここで、顧客 i のカテゴリ 1 のブランド b_j^1 からカテゴリ 2 のブランド b_k^2 へのブランド選択確率（これを $P_i(j,k)$ とする）を考えてみよう。すると、顧客 i においてその確率はカテゴリ 1 におけるブランドシェアと等しくなると考えられる。したがって、カテゴリ 1 のブランド b_j^1 の選択に依存することはなく、以下のようになる。

$$P_i(j,k) = a_i^2(k) / a_i^2 \tag{4}$$

次に通常の顧客が何らかの選好を持っているものとして、カテゴリ 1 のブランド b_j^1 からカテゴリ 2 のブランド b_k^2 へのブランド選択確率（これを P(j,k)とする）を考えてみよう。これは、ブランド b_j^1 の購入量で重み付けされた $P_i(j,k)$ の 2 乗誤差を最小化する値 x として定義することができ、例えば x は以下の問題を解くことで計算される。

$$\min \sum_{i=1}^{n} a_i^1(j)(P_i(j,k) - x)^2 \tag{5}$$

ここでも、購入頻度の高い顧客の購買行為が x の評価に強く反映しており、現実をうまく表現していると考えられる。このようにして計算された x は、カテゴリ 1 のブランド b_j^1 からカテゴリ 2 のブランド b_k^2 へのブランドスウィッチ確率を表している。この値とカテゴリ 2 におけるブランド b_k^2 のシェアとの相対的大きさが、カテゴリ 1 のブランド b_j^1 との相関の高さを表している。$as(j,k)$ で表されるブランド b_j^1 からカテゴリ 2 のブランド b_k^2 への

association strength は、カテゴリ C_2 の b_k^2 のブランドシェアを $share^1(k)$ とすると、次のように定義できる。

$$as(j,k) = P(j,k) / share^1(k) \quad (6)$$

我々は、異なる商品群に属する k 個のブランドの組み合わせにまで association strength の考え方を拡張すると、すべてのブランドの組み合わせにおける association strength の合計として単純に定義することができる。

結果

提案した association strength の有用性を示すために、日本のドラッグストアチェーンの顧客の購買履歴を利用 [4, 6, 8, 12]し、２つの実験を行った。

赤ちゃん用紙おむつに関する時系列ブランド関連分析

第一の実験の対象商品は赤ちゃん用紙おむつであり、おむつユーザーのブランドスウィッチ行為とブランドロイヤリティを議論する。まず、1996 年から 1999 年の 4 年間において、すべてのサイズを購入したオムツユーザー、3688 人をセレクトし実験を行った。

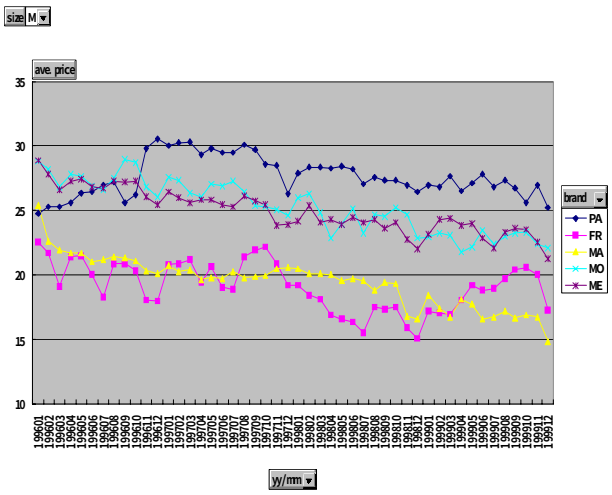


図 1　主要 5 ブランドの平均単価推移

日本における紙おむつのマーケットには、5 つの主要なブランドが存在する。これらのブランドの平均単価は図 1 の通りである。グラフから分かるように、近年、価格帯によってこれらのブランドは 3 つのグループに分類できる。グループ 1 は最も価格が高い C、最も価格の低いブランド D、E がグループ 2。そしてその中間の価格帯であるブランド A、B がグループ 3 である。これらのグラフから価格帯が同じブランド A と B、そしてブランド D と E が競合関係にあるものと推測できるし、業界でもそ

サイズ	A	B	C	D	E	その他
L	40.5%	16.4%	26.8%	16.0%		0.2%
M	32.2%	18.7%	23.3%	25.6%		0.2%
S	24.7%	22.1%	14.6%	18.5%	12.6%	7.4%
SS	21.8%	26.5%	11.5%	16.7%	14.6%	8.8%

表 2 主要 5 ブランドのサイズ別シェア

のように考えられていた。

表 2 は、小児用の SS、S、M、L サイズそれぞれの主要 5 ブランドのマーケットシェアを表したものである。ブランド E は、SS サイズと S サイズを作っておらず、そしてブランド A と E は同じメーカーによって作られていた。サイズごとのブランドシェアには、かなりばらつきがあり、サイズが変わるたびに何らかのブランドスウィッチが起こっていることが推測できる。

我々の目的は一連のサイズ間におけるブランドスウィッチ行為を測定することである。例えば、ブランド A の紙おむつメーカーを考えてみよう。メーカーにとって、M サイズのブランド A を使用している顧客が L サイズのときにどのブランドにスウィッチするか、そして L サイズでブランド A を使い始めた顧客は M サイズのどのブランドからスウィッチしてきたかについての顧客知識を得ることは重要である。我々は association strength を用いてそのようなブランドスウィッチパターンを分類した。

ブランド	SS→S	S→M	M→L
A	2.21	1.92	1.91
B	2.88	2.65	2.27
C	2.76	3.49	5.53
D	2.76	2.51	2.74
E	-	-	3.80

表 3　サイズ間における同一ブランドの association strength

おむつのサイズは小児用の SS から S、M、L サイズまで４つあり、主要５ブランドを対象にした（ただし１ブランドだけ M サイズと L サイズしか持っていない）。

表３は、サイズ間における５ブランドのブランドロイヤリティを測定するために、継続購入を示す同一ブランドへのassociation strength[1]を計算したものである。これによると、異なるサイズ間での同一ブランドへのassociation strengthは、他のどのブランドのサイズ間よりもブランドCが最も高い。これは、ブランドCがサイズを問わず高いブランドロイヤリティを維持していることを表しており、実際にCの平均価格帯は他の商品と比較して高いにもかかわらず、継続購入の割合が高かった。全体の傾向として、同一ブランドのサイズ間のassociation strengthは、異なるブランドよりも高くなる傾向にある。これは一般的な常識とも合致するもので、我々のアイデアの妥当性を示しているといえよう。

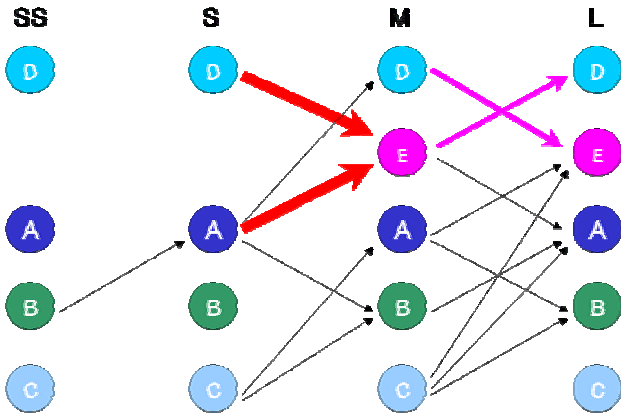


図 2　サイズ間における異なるブランドへのブランドスウィッチ

[1] 本稿では、同一ブランドの継続購入も同一ブランドへのスウィッチ行為と考え、ブランドスウィッチに広く含まれるものと考えている。

次にあるブランドから他のブランドへの association strength が比較的高いということは、ブランドスウィッチ率が高いということを示している。　図２は、サイズ間における異なるブランドへのブランドスウィッチ状況を表している。図の中では、association strength が１よりも大きいものを赤の太線、0.9～1.0 をピンクの中線、0.5 から 0.9 を点線で表現している。全体の傾向として、サイズが大きくなるにつれて、比較的高い価格帯であるブランド A や B から、低価格帯の D や E にブランドスウィッチしていることがわかる。

ここで注目すべきことは、S サイズから M サイズへの移行の際にブランド A からブランド E に顕著なブランドスウィッチが起こっている点である。実はこれらのブランドは同一のメーカーが製造するもので、カニバリゼーションが起こっていることを意味する。そもそもブランド E は､競合他社メーカーが出した低価格商品 D に対抗するために出されたもので､自社ブランド A のブランド力維持のために別ブランドとして差別化されたものであった。しかしながら、実際の顧客の購買行為はメーカーの思惑とは大きく異なり、競合ブランド D から顧客を奪うどころか､自社ブランド A の顧客の低価格帯へのブランドスウィッチを引き起こし、顧客一人あたりの売上単価の低下をもたらしていることが明らかになった。

複数カテゴリ間におけるブランドポジション分析

2 つめの実験では、複数カテゴリに属する複数ブランド間の購買関連性を対象に実験を行った。

メーカーにとって、自社のブランドがどの程度幅広くユーザーに認識されているのかを理解することは重要である。この観点から言うと、association strength を求めることで、他のメーカーのどのブランドが自社のブランドの強い購買関連性があるのかを理解することができ、それによって同一商品カテゴリにおける相対的なブランドポジションの知識を獲得することができる。

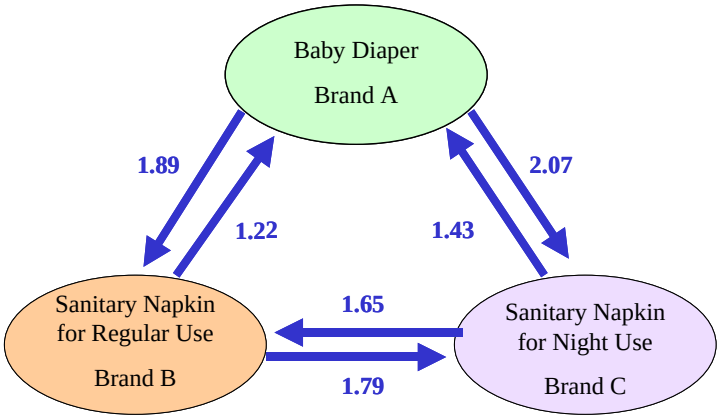
我々は上述の日本のドラッグストアチェーンで蓄積された POS データで実験を行った。我々の実験では 252,761 人の顧客データを対象とした。また商品カテゴリは同チェーンが採用している商品カテゴリ分類 750 カテゴリがあり、カテゴリあたりの平均ブランド数は約 15 であった。そのうち、売上シェアの高い 50 カテゴリを対象にし、それぞれのカテゴリ内で 10%以上のシェアを持つブランドに対象を限定した。

実験の結果、我々は以下のような興味深い結果を得ることができた。

（1）

我々は、カテゴリのすべての 3 つの組み合わせにおける、それぞれのカテゴリに属する 3 つのブランドのすべての組み合わせにおいて、association strength を求めた。得られた結果の中で最も高い値を出した３つのブランド群の 1 つを図 3 で表してある。カテゴリは赤ちゃん用紙おむつ、昼用生理用品、夜用生理用品の３つである。図３に出ているブランドは、その価格帯が比較的低いもの、そして頻繁に割引セールが行われる対象商品であっ

た。これらの顧客はそれぞれのブランドにロイヤリティを持っているのではなく、それらの価格にロイヤリティを持っていることが分かる。従来、いわゆるバーゲンハンターと呼ばれる顧客が存在することは認識されていたが、それらの顧客には商品群間の低価格商品に対する購買関連性があることは知られていなかった。



Total Score:10.05

図 3　最も購買関連性の高いブランド群

(1)	(2)	(3)	(4)	(5)	Total Score
A(a)	B(b)	E(c)	H(a)	K(a)	24.7
B(b)	D(b)	F(b)	I(b)	L(b)	24.2
C(c)	B(b)	G(c)	J(a)	M(a)	23.5

表 4　メーカーロイヤリティの例

(1)衣料用粉末洗剤、(2)液体衣料用洗剤、(3)柔軟剤、(4)食器用洗剤、(5)リンス・イン・シャンプー

（2）

我々は、同様の実験を 5 カテゴリで行い、表 4 の結果を得た。括弧内の a、b、c はそれぞれのブランドのメーカーを表している。5 つのカテゴリの組み合わせすべてからすべてのブランドの組み合わせで高い値が 3 つ見つかった。表 4 から分かるように、2 つめの組み合わせは、メーカーb のすべてのブランドに対してロイヤリティを持っている顧客群が存在していることを示している。他の 2 つの組み合わせは、液体洗剤のカテゴリにおいてメーカーa と b がそれほど強くないことを示しており、柔軟剤ではメーカーc、そして食器用洗剤とリンス in シャンプーではメーカーa がロイヤリティの高い顧客群をもっていることが分かる。特に競合関係の強いこの 3 メーカーにとって、これらの情報は自社の競争ポジションを理解するのに有用である。

考察

本稿では、複数購入期間における複数ブランド間の購買関連性を測定する association strength の考え方を提示した。そしてドラッグストアの POS データに本手法を適用することによって、現実のビジネスに対して重要な知見を得ることができた。これらの手法は、複数カテゴリの複数ブランド間の競合関係にも拡張することができ、従来のアソシエーションルールでは扱えなかった競合分析の領域を大きく拡張できる可能性を示すことができたと考えられる。

ただし、これらの知見はあくまで、現在の競争状況を表現しているに過ぎず、それらを打開していく能動的なビジネスアクションを与えるものではない。ビジネスアクションは企業戦略や商品戦略を基盤にして形成されるもので、時と場合によって様々なバリエーションが必要である。本稿で提示した考え方も、これらの必要なビジネスアクションにあわせてさらに改良されていくべきであろう。

来年度以降は、これらの成果に基づき、様々なビジネス領域（食品、アパレルなど）の様々な競争状態のブランド群に適用、問題点を明らかし、詳細な検討を加える。また、時間軸を２～３年という比較的長期間なものから１～２週間という短期間なものまで、多様な問題に当てはめ、時系列データの取り扱いに関する統一的な枠組みを検討する。

参考文献

[1] R. Agrawal, T. Imielinski and A. Swami, Database mining: A performance perspective, IEEE Transactions on Knowledge and Data Engineering, 5, pp.914-925 (1993).

[2] K. Fujisawa, Y. Hamuro, N. Katoh, T. Tokuyama and K. Yada, Approximation of Optimail Two-Dimensional Association Rules for Categorical Attributes Using Semidefinite Programming, Proc. of 2nd Symp. on Discovery Science, LNAI 1721, Springer-Verlag, pp.148-159 (1999).

[3] Y. Hamuro, E. Kawata, N. Katoh and K. Yada, A Machine Learning Algorithm for Analyzing String Patterns Helps to Discover Simple and Interpretable Business Rules from Purchase History, to appear in Progresses in Discovery Science, State-of-the-Art Surveys, LNCS, Springer-Verlag (2001).

[4] Y. Hamuro, N. Katoh, Y. Matsuda and K. Yada, Mining Pharmacy Data Helps to Make Profits, Data Mining and Knowledge Discovery, Vol.2, No.4, pp.391-398 (1998).

[5] Y. Hamuro, N. Katoh and K. Yada, Discovering Interpretable Rules that Explain Customer Brand Choice Behavior, INFORMS-KORMS Seoul 2000, pp.561-568 (2000).

[6] Y. Hamuro, N. Katoh and K. Yada, Discovering Association Strength among Brand Loyalties from Purchase History, Proceeding of 2001 IEEE International Symp. on

Industrial Electronics, pp.114-117 (2001).
[7] E. Ip, K. Yada, Y. Hamuro, and N. Katoh, A Data Mining System for Managing Customer Relationship, Proc. of the 2000 Americas Conference on Information Systems, pp.101-105 (2000).
[8] 加藤直樹, 羽室行信, 矢田勝俊: 新規顧客からのロイヤルカスタマーの早期発見, ESTRELA, No.89, pp.10-17 (2001).
[9] B. Kitts, D. Freed and M. Vrieze, Cross-sell: A Fast Promotion-Tunable Customer-item Recommendation Method Based on Conditionally Independent Probabilities, Proc. KDD 2000, pp.437-446 (2000).
[10] G. Paitetsky-Shapiro (Editor), Knowledge Discovery in Databases, AAAI Press (1991).
[11] R. Uthurusamy, U. M. Fayyad and S. Spangler, Learning Useful Rules from Inconclusive Data, In [10], pp.141-157 (1991).
[12] K. Yada, Y. Hamuro and N. Katoh, The Discovery of Customer Loyalty from Newcomers, Proc. of 2000 MIS/OA International Conference, pp.185-189 (2000).
[13] 矢田勝俊, 羽室行信, 加藤直樹: 経営データからの知識発見, 国民経済雑誌, 第 184 巻第 1 号, pp.19-33 (2001).
[14] 矢田勝俊, 飯田洋: ビジネス知識発見の研究展望とアクティブマイニング, 人工知能基礎論研究会, SIG-FAI/KBS-J-35, pp.213-217 (2001).

Document Clustering by a Tolerance Rough Set Model

Head of group	Tu Bao Ho	(Japan Adv. Inst. of Science and Technology)
Collaborator	Saori Kawasaki	(Japan Adv. Inst. of Science and Technology)
Collaborator	Ngoc Binh Nguyen	(Hanoi University of Technology)

1 Introduction

Document clustering, the grouping of documents into several clusters, has been recognized as a means of improving efficiency and effectiveness of text retrieval. With the growing importance of electronic media for storing and exchanging large textual databases, document clustering becomes more significant. Document clustering helps the user to exploit large document collections in several ways, such as it enables the user to select and tackle only part of the collection that is relevant to his/her interest, or it assists the user from members and representatives of clusters to uncover topics, hypotheses, concepts, or novel nuggets. However, document clustering is a difficult clustering problem by a number of reasons [2], [4], [14]. The main difficulty comes from the unstructured form and textual characteristics of documents. As a consequence, the quality of document clustering not only depends on clustering algorithms but also largely depends on document representation models.

Rough set theory, a mathematical tool to deal with vagueness and uncertainty introduced by Pawlak in early 1980s [7], has been successful in many applications [5], [8]. In this theory each set in a universe is described by a pair of ordinary sets called lower and upper approximations, determined by an equivalence relation in the universe. The use of the original rough set model in information retrieval, called the *equivalence rough set model* (ERSM), has been investigated by several researchers [9], [13]. A significant contribution of ERSM to information retrieval is that it suggested a new way to calculate the semantic relationship of words based on an organization of the vocabulary into equivalence classes. However, as analyzed in [3], ERSM is not suitable for information retrieval and text processing in general due to the fact that the requirement of the transitive property in equivalence relations is too strict to the meaning of words, and there is no way to calculate automatically equivalence classes of terms. Inspired by some works that employs different relations to generalize new models of rough set theory, e.g., [11], [12], a *tolerance rough set model* (TRSM) for text processing that adopts tolerance classes instead of equivalence classes has been developed [3].

In this work we extend TRSM in [3] and introduce two TRSM-based hierarchical and nonhierarchical document clustering algorithms, as well methods for cluster-based information retrieval (IR). These algorithms have been evaluated and vali-

dated experimentally on IR test collections. The results show advantages of the model, particularly in improving precision in information retrieval.

Section 2 of the paper presents the TRSM for representing documents. Section 3 describes two TRSM clustering algorithms. Section 4 presents a evaluation and validation of these algorithms, and section 5 addresses the TRSM cluster-based information retrieval.

2 The tolerance rough set model

Given a set $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ of M full text documents. Each document d_j is mapped into a list of terms t_i each is assigned a weight that reflects its importance in the document. Denote by $f_{d_j}(t_i)$ the number of occurrences of term t_i in d_j (term frequency), and by $f_{\mathcal{D}}(t_i)$ the number of documents in $\mathcal{D}$ that term t_i occurs in (document frequency). The weights w_{ij} of terms t_i in documents d_j are first calculated by

$$w_{ij} = \begin{cases} (1 + \log(f_{d_j}(t_i))) \times \log \frac{M}{f_{\mathcal{D}}(t_i)} & \text{if } t_i \in d_j, \\ 0 & \text{if } t_i \notin d_j \end{cases} \tag{1}$$

then normalized by $w_{ij} \leftarrow w_{ij}/\sqrt{\sum_{t_{hj} \in d_j} (w_{hj})^2}$. Each document d_j is represented by its r highest-weighted terms, i. e., $d_j = (t_{1j}, w_{1j}; t_{2j}, w_{2j}; \ldots; t_{rj}, w_{rj})$ where $w_{ij} \in [0, 1]$. A usual way is to fix a default value r common for all documents. The set of all terms from $\mathcal{D}$ and queries Q are denoted with $q_i \in \mathcal{T}$ and $w_{iq} \in [0, 1]$ by

$$\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$$
$$Q = (q_1, w_{1q}; q_2, w_{2q}; \ldots; q_s, w_{sq})$$

The *tolerance rough set model* (TRSM) aims to enrich the document representation in terms of semantics relatedness by creating tolerance classes of terms in $\mathcal{T}$ and approximations of subsets of documents. The model has the root from rough set models and its extensions [7], [11]. The key idea is among three properties of an equivalence relation R in an universe U used in the original rough set model (reflexive: xRx; symmetric: $xRy \rightarrow yRx$; transitive: $xRy \wedge yRz \rightarrow xRz$ for $\forall x, y, z \in U$), the transitive property does not always hold in natural language processing, information retrieval, and consequently text data mining. In fact, words are better viewed as overlapping classes which can be generated by *tolerance relations* (requiring only reflexive and symmetric properties).

Denote by $f_{\mathcal{D}}(t_i, t_j)$ the number of documents in $\mathcal{D}$ in which two index terms t_i and t_j co-occur. We define an uncertainty function I depending on a threshold θ:

$$I_\theta(t_i) = \{t_j \mid f_{\mathcal{D}}(t_i, t_j) \geq \theta\} \cup \{t_i\} \tag{2}$$

It is clear that the function I_θ defined above satisfies the condition of $t_i \in I_\theta(t_i)$ and $t_j \in I_\theta(t_i)$ iff $t_i \in I_\theta(t_j)$ for any $t_i, t_j \in \mathcal{T}$, and so I_θ is both reflexive and

symmetric. This function corresponds to a tolerance relation $\mathcal{I} \subseteq \mathcal{T} \times \mathcal{T}$ that $t_i \mathcal{I} t_j$ iff $t_j \in I_\theta(t_i)$, and $I_\theta(t_i)$ is the tolerance class of index term t_i. A vague inclusion function ν, which determines how much X is included in Y, is defined as

$$\nu(X, Y) = \frac{|X \cap Y|}{|X|} \tag{3}$$

This function is clearly monotonous with respect to the second argument. Using this function the membership function μ for $t_i \in \mathcal{T}, X \subseteq \mathcal{T}$ can be defined as

$$\mu(t_i, X) = \nu(I_\theta(t_i), X) = \frac{|I_\theta(t_i) \cap X|}{|I_\theta(t_i)|} \tag{4}$$

With these definitions we can define a tolerance space as $\mathcal{R} = (\mathcal{T}, I, \nu, P)$ in which the *lower approximation* $\mathcal{L}(\mathcal{R}, X)$ and the *upper approximation* $\mathcal{U}(\mathcal{R}, X)$ in $\mathcal{R}$ of any subset $X \subseteq \mathcal{T}$ can be defined as

$$\mathcal{L}(\mathcal{R}, X) = \{t_i \in \mathcal{T} \mid \nu(I_\theta(t_i), X) = 1\} \tag{5}$$

$$\mathcal{U}(\mathcal{R}, X) = \{t_i \in \mathcal{T} \mid \nu(I_\theta(t_i), X) > 0\} \tag{6}$$

The vector length normalization is then applied to the upper approximation $\mathcal{U}(\mathcal{R}, d_j)$ of d_j. Note that the normalization is done when considering a given set of index terms.

The term-weighting method defined by Eq. (1) is extended to define weights for terms in the upper approximation $\mathcal{U}(\mathcal{R}, d_j)$ of d_j. It ensures that each term in the upper approximation of d_j but not in d_j has a weight smaller than the weight of any term in d_j.

$$w_{ij} = \begin{cases} (1 + \log(f_{d_j}(t_i))) \times \log \frac{M}{f_{\mathcal{D}}(t_i)} & t_i \in d_j, \\ \min_{t_{hj} \in d_j} w_{hj} \times \frac{\log(M/f_{\mathcal{D}}(t_i))}{1+\log(M/f_{\mathcal{D}}(t_i))} & t_i \in \mathcal{U}(\mathcal{R}, d_j) \setminus d_j \\ 0 & t_i \notin \mathcal{U}(\mathcal{R}, d_j) \end{cases} \tag{7}$$

3 TRSM Clustering Algorithms

3.1 Algorithms

Table 2 describes the general TRSM-based hierarchical clustering algorithm that is an extension of the hierarchical agglomerative clustering algorithm. The main point here is at each merging step it uses upper approximations of documents in finding two closest clusters to merge. Several variants of agglomerative clustering can be applied, such single-link or complete-link clustering.

As documents are represented as length-normalized vectors and when cosine similarity measure is used, an efficient alternative is to employ the group-average agglomerative clustering. The group-average clustering avoids the elongated and

MED_1: correlation between maternal and fetal plasma levels of glucose and free fatty acids . correlation coefficients have been determined between the levels of glucose and ffa in maternal and fetal plasma collected at delivery . significant correlations were obtained between the maternal and fetal glucose levels and the maternal and fetal ffa levels . from the size of the correlation coefficients and the slopes of regression lines it appears that the fetal plasma glucose level at delivery is very strongly dependent upon the maternal level whereas the fetal ffa level at delivery is only slightly dependent upon the maternal level .

MED_1: 21-0.178679, 44-0.094230, 48-0.228942, 57-0.235588, 110-0.257558, 198-0.328567, 299-0.126899, 403-0.371317, 437-0.136658, 683-0.306114, 692-0.306114, 694-0.306114, 1840-0.289422, 2546-0.189904, 4546-0.321535.

Table 1: A document and its TRSM representation

straggling clusters produced by single-link clustering, and avoids the high cost of complete link clustering. In fact, it allows using cluster representatives to calculate the similarity between two clusters instead of averaging similarities of all document pairs each belong to one cluster [6], [14]. In such a case, the complexity of computing average similarity would be $O(N^2)$.

Table 3 describes the TRSM nonhierarchical clustering algorithm. It can be considered as a reallocation clustering method to form K clusters of a collection $\mathcal{D}$ of M documents [2].

The distinction of the TRSM nonhierarchical clustering algorithm is it forms overlapping clusters and it uses approximations of documents and cluster's representatives in calculating their similarity. The latter allows us to find some semantic relatedness between documents even when they do not share common index terms. After determining initial cluster representatives in step 1, the algorithm mainly consists of two phases. The first does an iterative reallocation of documents into overlapping clusters by steps 2, 3 and 4. The second does by step 5 an assignment of documents that are not classified in the first phase, into clusters containing their nearest neighbors with non-zero similarity.

We consider two other issues that have an important influence on the clustering quality (i) how to define the representatives of clusters; and (ii) how to determine the similarity between documents and the cluster representatives.

3.2 Representatives of clusters

The TRSM clustering algorithm constructs a *polythetic* representative R_k for each cluster $C_k, k = 1, \ldots, K$. In fact, R_k is a set of index terms such that:

(i) each document $d_j \in C_k$ has some or many terms in common with R_k;
(ii) terms in R_k are possessed by a large number of $d_j \in C_k$;
(iii) no term in R_k must be possessed by every document in C_k.

Input $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$
a similarity measure $\mathcal{S} : \mathcal{P}(\mathcal{D}) \times \mathcal{P}(\mathcal{D}) \rightarrow R^+$
Result A hierarchical structure of $\mathcal{D}$

1. Initially, consider each document of $\mathcal{D}$ as a cluster with one member $C_j = \{d_j\}$, and $H = \{C_1, C_2, , \ldots, C_M\}$.
2. Identify two most similar clusters in terms of upper approximations of their representatives, $(C_{n_1}, C_{n_2}) = argmax_{(C_u, C_v) \in H \times H} \mathcal{S}(\mathcal{U}(\mathcal{R}, C_u), \mathcal{U}(\mathcal{R}, C_v)$
3. Form a new cluster $C_i = C_{n_1} \cup C_{n_2}$ and let $H = (H \setminus \{C_{n_1}, C_{n_2}\}) \cup \{C_i\}$.
4. If more than one cluster remains, return to steps 2 and 3.

Table 2: TRSM hierarchical clustering algorithm

It is known that the Bayesian decision rule with minimum error rate will assign a document d_j in the cluster C_k if

$$P(d_j|C_k)P(C_k) > P(d_j|C_h)P(C_h), \forall h \neq k \tag{8}$$

With the assumption that terms occur independently in documents, we have

$$P(d_j|C_k) = P(t_{j_1}|C_k)P(t_{j_2}|C_k) \ldots P(t_{j_p}|C_k) \tag{9}$$

Denote by $f_{C_k}(t_i)$ the number of documents in C_k that contain t_i, we have $P(t_i|C_k) = f_{C_k}(t_i)/|C_k|$. Equation (9) and heuristics of the polythetic properties of the cluster representatives lead us to adopt rules to form the cluster representatives:

(i) Initially, $R_k = \phi$.
(ii) For all $d_j \in C_k$ and for all $t_i \in d_j$, if $f_{C_k}(t_i)/|C_k| > \sigma$ then $R_k = R_k \cup \{t_i\}$.
(iii) If $d_j \in C_k$ and $d_j \cap R_k = \phi$ then $R_k = R_k \cup \text{argmax}_{t_i \in d_j} w_{ij}$.

In case of group-average clustering, σ could be 0 to ensure the use of cluster representatives when calculating the cluster similarity. The weights of terms t_i in R_k is first averaged by of weights of this terms in all documents belonging to C_k, that means $w_{ik} = (\sum_{d_j \in C_k} w_{ij})/|\{d_j : t_i \in d_j\}|$, then normalized by the length of the representative R_k.

3.3 Document and cluster similarity

Many similarity measures between documents can be used in the TRSM clustering algorithm. Three common coefficients of Dice, Jaccard and Cosine [2] are implemented in the TRSM clustering program to calculate the similarity between pairs

Input $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ and the number K.
Result K clusters of documents with their membership.

1. Determine the initial representatives $R_1, R_2, ..., R_K$ of clusters $C_1, C_2, ..., C_K$ as K randomly selected documents in $\mathcal{D}$.
2. For each $d_j \in \mathcal{D}$, calculate the similarity $S(\mathcal{U}(\mathcal{R}, d_j), R_k)$ between its upper approximation $\mathcal{U}(\mathcal{R}, d_j)$ and the cluster representative $R_k, k = 1, ..., K$. If this similarity is greater than a given threshold, assign d_j to C_k and take this similarity value as the cluster membership $m(d_j)$ of d_j in C_k.
3. For each cluster C_k, re-determine its representative R_k.
4. Repeat steps 2 and 3 until there is little or no change in cluster membership during a pass through $\mathcal{D}$.
5. Denote by d_u an unclassified document after steps 2, 3, 4 and by $\mathrm{NN}(d_u)$ its nearest neighbor document (with non-zero similarity) in formed clusters. Assign d_u into the cluster that contains $\mathrm{NN}(d_u)$, and determine the cluster membership of d_u in this cluster as the product $m(d_u) = m(\mathrm{NN}(d_u)) \times S(\mathcal{U}(\mathcal{R}, d_u), \mathcal{U}(\mathcal{R}, \mathrm{NN}(d_u)))$. Re-determine the representatives R_k, for $k = 1, ..., K$.

Table 3: TRSM nonhierarchical clustering algorithm

of documents d_{j_1} and d_{j_2}. For example, the cosine coefficient is

$$S_C(d_{j_1}, d_{j_2}) = \frac{\sum_{k=1}^{N}(w_{kj_1} \times w_{kj_2})}{\sqrt{\sum_{k=1}^{N} w_{kj_1} \times \sum_{k=1}^{N} w_{kj_2}}} \tag{10}$$

It is worth to note that the cosine coefficient (or any other well-known similarity coefficient used for documents [2]) yields a large number of zero values when documents are represented by r terms as many of them may have no terms in common. The use of the tolerance upper approximation of documents and of the cluster representatives allows the TRSM algorithm to improve this situation. In fact, in the TRSM clustering algorithm, the normalized cosine coefficient is applied to the upper approximation of documents $\mathcal{U}(\mathcal{R}, d_j)$ and cluster representatives $\mathcal{U}(\mathcal{R}, R_k)$. Two main advantages of using upper approximations are: (i) To reduce the number of zero-valued coefficients by considering documents themselves together with the related terms in tolerance classes; and (ii) The upper approximations formed by tolerance classes make it possible to relate documents that may have few (even no) terms in common with the user's topic of interest or the query.

Collection	Subject	doc.	query	rel. doc.
JSAI	AI	802	20	32
CACM	Comp. Sci.	3200	64	15
CISI	Lib. Sci.	1460	76	40
CRAN	Aero.	1400	225	8
MED	Medicine	3078	30	23

Table 4: Test collections

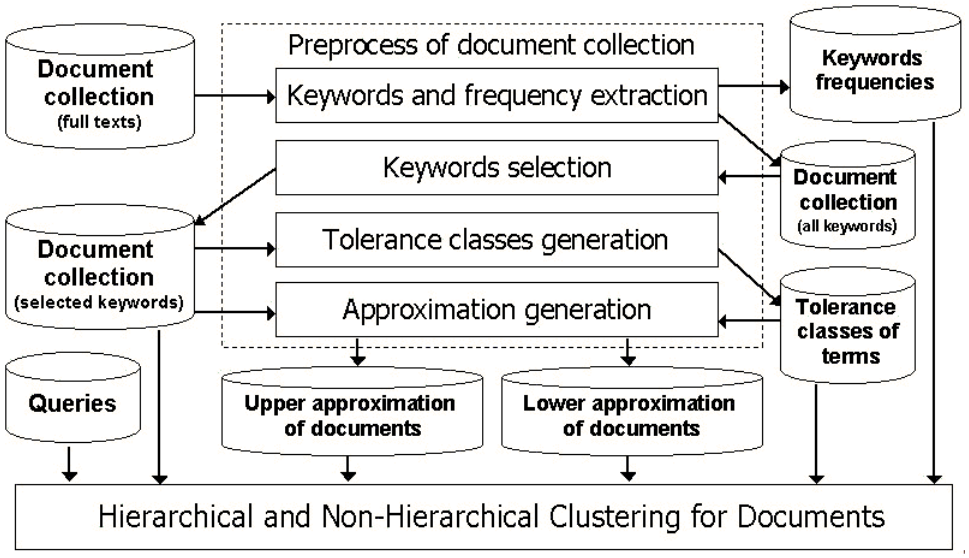


Figure 1: Conceptual architecture of the system

4 Validation and Evaluation

Figure 1 shows the conceptual architecture of the TRSM clustering system. Table 4 summarizes test collections used in our experiments including JSAI where each document is represented in average by 5 keywords and four other common test collections CACM, CISI, CRAN and MED [2]. Columns 3, 4, and 5 show the number of documents, queries, and the average numbers of relevant documents for queries. The clustering quality for each test collection depends on parameter θ in TRSM and on σ in clustering algorithm. We can note that the higher value of θ the large upper approximation and the smaller lower approximation of a set X. Our experiments suggested that when the average number of terms in documents is high and/or the size of the document collection is large, high values of θ are often appropriate and vice-versa. In Table 9 we can see how retrieval effectiveness relates to different values of θ. To avoid biased experiments when comparing algorithms we

	% average of relevant documents						
	0	1	2	3	4	5	avg.
JSAI	19.9	19.8	18.5	18.5	11.8	11.5	2.2
CACM	50.3	22.5	12.8	7.9	4.2	2.3	1.0
CISI	45.4	25.8	15.0	7.5	4.3	1.9	1.1
CRAN	33.4	32.7	19.2	9.0	4.6	1.0	1.2
MED	10.4	18.7	18.6	21.6	19.6	11.1	2.5

Table 5: Results of clustering tendency

take default values $\theta = 15$, and $\sigma = 0.1$ for all five test collections.

4.1 Validation of Clustering Tendency

The experiments for clustering tendency "attempt to determine whether worthwhile retrieval performance would be achieved by clustering a document collection, before investing the computational resources which clustering the database would entail" [2]. We employ the *nearest neighbor test* [14] by considering, for each relevant document of a query, how many of its n nearest neighbors are also relevant; and by averaging over all relevant documents for all queries in a test collection in order to obtain single indicators. We use in these experiments five test collections with all queries and their relevant documents.

The experiments are carried out to calculate the percentage of relevant documents in the database that had 0, 1, 2, 3, 4, or 5 relevant documents in the set of 5 nearest neighbors of each relevant document. Table 5 reports the experimental results synthesized from those done on five test collections. Columns 4 and 5 show the number of queries and total number of relevant documents for all queries in each test collection. The next six rows stand for the percentage average of the relevant documents in a collection that had 0, 1, 2, 3, 4, and 5 relevant documents in their sets of 5 nearest neighbors. For example, the meaning of row JSAI column 11 is "among all relevant documents for 20 queries of JSAI collection, 11.5 % of them have 5 nearest neighbor documents are all relevant documents". The last column shows the average number of relevant documents among 5 nearest neighbors of each relevant document. This value is relatively high for JSAI and MED collections and vice-versa for the others.

As the finding of nearest neighbors of a document in this method is based on the similarity between the upper approximations of documents, this tendency suggests if the TRSM clustering method might appropriately be applied for retrieval purpose. This tendency can be clearly observed in concordance with the high retrieval effectiveness for JSAI and MED shown in Table 9.

θ	Percentage of changed data						
	1%	2%	3%	4%	5%	10%	15%
2	2.84	5.62	7.20	5.66	5.48	11.26	14.41
3	3.55	4.64	4.51	6.33	7.93	12.06	15.85
4	0.97	2.65	2.74	4.22	5.62	8.02	13.78

Table 6: Synthesized results about the stability

4.2 Validation of Clustering Stability

The experiments were done for the JSAI test collection in order to validate the stability of the TRSM clustering, i.e., to verify that whether the TRSM clustering method produces a hierarchy which is unlikely to be altered drastically when further documents are incorporated. For each value 2, 3, and 4 of θ, the experiments are done 10 times each for a reduced database of size $(100 - s)\%$ of $\mathcal{D}$. We randomly removed a specified amount of $s\%$ documents from the JSAI database, then re-determine the new tolerance space for the reduced database. Once having the new tolerance space, we perform the TRSM clustering algorithm and evaluate the change of clusters due to the change of the database. Table 6 synthesizes the experimental results with different values of s from 210 experiments with $s\% = 1\%$, 2%, 3%, 4%, 5%, 10% and 15%.

Note that a little change of data implies a possible little change of hierarchy (about at the same percentage as for $\theta = 4$). The experiments on the stability for other test collections have nearly the same results as those of JSAI. It suggests that the TRSM hierarchical clustering results are stable.

4.3 Hierarchical Clustering Efficiency

From a given collection of documents we need to prepare all the files before running the TRSM clustering algorithms. It consists of making an index term file, term encoding, document-term and term-document (inverted) relation files as indexing files, files of term co-occurrences and tolerance classes for each value of θ. A direct implementation of these procedures requires the time complexity of $O(M + N^2)$, but we implemented the system by applying a sorting algorithm (quick-sort) of $O(N \log N)$ to make the indexing files, then created the TRSM related files for the term co-occurrence, tolerance classes, upper and lower approximations files in the time of $O(M + N)$.

The experiments reported in this paper were performed on a conventional workstation GP7000S Model 45 (Fujitsu, 250 MHz Ultra SPARC-II, 512 MB). We can note that the search for clusters requires in average $\log M$, then the search will be done with a subset of documents in the clusters. However, the time complexity of the clustering is of $O(M^2 + N)$, and the space is of $O(M^2 + N)$ because of using an

Col.	Size (MB)	Nb. of doc.	Nb. of query	TRSM Time	HC Time	NHC Time	Full Search	1-Cluster Search	HM (MB)	NHM (MB)
JSAI	0.1	802	20	2.4s	14.9s	8.0s	0.8s	0.1s	8	12
CACM	2.2	3200	64	22m2.2s	26m46.8s	2m26s	13.3s	1.2s	201	15
CISI	2.2	1460	76	13m16.8s	4m49.8s	18s	40.1s	3.4s	84	13
CRAN	1.6	1400	225	23m9.9s	3m6.9s	13s	20.5s	1.8s	71	13
MED	1.1	1033	430	0.1s	1m30.8s	4s	2.5s	0.3s	25	28

Table 7: Performance Measurements of the TRSM Cluster-based Retrieval

	1.2% (0.18)		1.8% (0.16)		2.9% (0.14)		8.0% (0.11)		16.9% (0.09)		full search	
Col.	P	R	P	R	P	R	P	R	P	R	P	R
JSAI	0.950	0.472	0.948	0.485	0.949	0.502	0.939	0.541	0.938	0.559	0.934	0.560
CACM	0.048	0.037	0.096	0.068	0.100	0.084	0.116	0.194	0.105	0.262	0.160	0.241
CISI	0.181	0.043	0.180	0.061	0.180	0.089	0.130	0.183	0.112	0.261	0.155	0.204
CRAN	0.121	0.127	0.140	0.149	0.139	0.173	0.139	0.214	0.112	0.245	0.257	0.301
MED	0.477	0.288	0.530	0.324	0.565	0.375	0.518	0.460	0.422	0.531	0.415	0.421
	1 cluster		2 clusters		3 clusters		4 clusters		5 clusters		full search	
Col.	P	R	P	R	P	R	P	R	P	R	P	R
JSAI	0.973	0.375	0.950	0.458	0.937	0.519	0.936	0.544	0.932	0.534	0.934	0.560
CACM	0.098	0.063	0.100	0.127	0.117	0.166	0.132	0.221	0.144	0.240	0.160	0.241
CISI	0.177	0.078	0.141	0.139	0.151	0.179	0.156	0.206	0.158	0.212	0.155	0.204
CRAN	0.204	0.219	0.238	0.278	0.250	0.290	0.257	0.301	0.261	0.304	0.257	0.301
MED	0.393	0.277	0.396	0.393	0.372	0.425	0.367	0.445	0.380	0.472	0.415	0.421

Table 8: Precision and recall of TRSM cluster-based retrieval and full retrieval

$M \times M$-matrix to store the similarities/distances between clusters in the hierarchy. Concerned with generating the TRSM files for the JSAI database, the direct implementation with $O(M + N^2)$ required up to 6 minutes [14 hours for CRAN], but the quicksort-based implementation with $O(N \log N)$ took about 3 seconds (JSAI) [23 minutes for CRAN] for making the files by running a package of shell scripts on UNIX. Table 7 summaries the time for generating the TRSM files, clustering, full search, cluster-based search, and the required memory size for each collection. The clustering time included the time for reading the TRSM files into the RAM memory. Thanks to short time for preparing the database files as well as shorter time for cluster-based search in comparing with the full search, the TRSM-based proposed method is able to be applied to large document collections.

4.4 TRSM Cluster-based Information Retrieval

We show the potential of the method in terms of cluster-based effectiveness and efficiency in application to information retrieval [2], [14]. The quality of generated hierarchy is evaluated in terms of information retrieval. The experiments evaluate effectiveness of the TRSM cluster-based retrieval by comparing it with full retrieval by using the common measures of *precision* and *recall*. Precision P is the ratio of the number of relevant documents retrieved over the total number of documents retrieved. Recall R is the ratio of relevant documents retrieved for a given query over the number of relevant documents for that query in the database. Precision and recall are defined as

θ	JSAI		CACM		CISI		CRAN		MED	
	P	R	P	R	P	R	P	R	P	R
30	0.934	0.560	0.146	0.231	0.147	0.192	0.265	0.306	0.416	0.426
25	0.934	0.560	0.158	0.242	0.151	0.194	0.266	0.310	0.416	0.426
20	0.934	0.560	0.159	0.243	0.150	0.194	0.268	0.311	0.416	0.426
15	0.934	0.560	0.160	0.241	0.155	0.204	0.257	0.301	0.415	0.421
10	0.934	0.560	0.141	0.221	0.142	0.178	0.255	0.302	0.414	0.387
8	0.934	0.560	0.151	0.254	0.138	0.172	0.242	0.291	0.393	0.386
6	0.945	0.550	0.141	0.223	0.146	0.178	0.233	0.271	0.376	0.365
4	0.904	0.509	0.137	0.182	0.152	0.145	0.223	0.241	0.356	0.383
2	0.803	0.522	0.111	0.097	0.125	0.057	0.247	0.210	0.360	0.193
VSM	0.934	0.560	0.147	0.232	0.139	0.184	0.258	0.295	0.429	0.444

Table 9: Precision and recall of TRSM and VSM full retrieval

$$P = \frac{|Rel \cap Ret|}{|Ret|} \qquad R = \frac{|Rel \cap Ret|}{|Rel|} \tag{11}$$

where $Rel \subset \mathcal{D}$ is the set of relevant documents in the database for the query, and $Ret \subset \mathcal{D}$ is the set of retrieved documents. Table 9 shows precision and recall of the TRSM-based full retrieval and the VSM-based full retrieval (Vector Space Model) where the TRSM-based retrieval is done with values 30, 25, 20, 15, 10, 8, 6, 4, and 2 of θ.

After ranking all documents according to the query, precision and recall are evaluated on the set of retrieved documents determined by the default *cutoff* value as the average number of relevant documents for queries in each test collection. From Table 9 we see that precision and recall for JSAI are high, and they are higher and stable for the other collections with $\theta \geq 15$. With these values of θ, the TRSM-based retrieval effectiveness is comparable or somehow higher than that of VSM.

4.4.1 Hierarchical cluster-based IR

We carried out retrieval experiments on all queries of test collections. Each query in the test collection is matched against the hierarchy from the root in the top-down direction in order to determine a subset $\mathcal{D}' \subset \mathcal{D}$. The subset $\mathcal{D}'$ is union of all clusters each has the similarity between the query and its representative greater than a threshold γ. The cluster-based retrieval is carried out in $\mathcal{D}'$.

Table 9 reports the average of precision and recall for all queries in test collections using the TRSM cluster-based retrieval with various proportion (%) of $|\mathcal{D}'|$ to $|\mathcal{D}|$), and full retrieval in whole $\mathcal{D}$ (accordingly, values of γ). The results show that in several cases (JSAI, CISI, and MED) just searching a small part of $\mathcal{D}$, says 1.2% or 1.8%, TRSM cluster-based search reaches precision higher than that of full search. Also, the TRSM cluster-based search achieved recall higher than that of full retrieval on most collections when $|\mathcal{D}'|$ is about 17% of $|\mathcal{D}|$.

Table **??** reports the effectiveness of TRSM cluster-based retrieval (TRSM) versus VSM cluster-based retrieval (VSM) when $|\mathcal{D}'|$ is 2.9%, 8.0%, and 16.9% of $|\mathcal{D}|$. It shows that TRSM cluster-based retrieval often achieves precision higher than

that of VSM cluster-based retrieval thought its recall is somehow lower. The results suggest that TRSM can be used to improve precision of information retrieval, and so in a certain tasks of text mining.

4.4.2 Nonhierarchical cluster-based IR

The lower half of Table 8 reports the average of precision and recall for all queries in test collections using the TRSM cluster-based retrieval with 1, 2, 3, 4 clusters, and full retrieval (15 clusters). Usually, along the ranking order of clusters when cluster-based retrieval is carried out on the more clusters we obtain higher recall value. Interestingly, the TRSM cluster-based retrieval achieved higher recall than that of full retrieval on several collections.

More importantly, the TRSM cluster-based retrieval on four clusters offers precision higher than that of full retrieval in most collections. Also, the TRSM cluster-based retrieval achieved recall and precision nearly as that of full search just after searching on one or two clusters. These results show that the TRSM cluster-based retrieval can contribute considerably to the problem of improving retrieval effectiveness in information retrieval.

5 Conclusion

We have proposed document hierarchical and non-hierarchical clustering algorithms based on the tolerance rough set model (TRSM) of tolerance classes of index terms, and developed a TRSM cluster-based method for information retrieval. Careful experiments have been done on test collections for evaluating and validating the proposed method on the clustering tendency and stability, the efficiency as well as effectiveness of cluster-based retrieval using the clustering results.

There are still many further works to do in this research: (1) to investigate the parameters of TRSM and their influence on text mining algorithms; (2) to incrementally update tolerance classes of terms and document clusters when new documents are added to the collections; (3) to extend the tolerance rough set model by considering the model without requiring a symmetric similarity or tolerance classes based on co-occurrence between more than two terms; and (4) to combine TRSM-based hierarchical and nonhierarchical clustering for very large text collections.

References

[1] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison Wesley, 1999.

[2] Fakes, W. B. and Baeza-Yates, *Information Retrieval. Data Structures and Algorithms* (eds.), Prentice Hall, 1992.

[3] Ho, T. B. and Funakoshi K., "Information retrieval using rough sets", *Journal of Japanese Society for Artificial Intelligence*, Vol. 13, N. 3, pp. 424–433, 1998.

[4] Lebart, L., Salem, A., and Berry, L., *Exploring Textual Data*, Kluwer Academic Publishers, 1998.

[5] Lin, T. Y. and Cercone, N., *Rough Sets and Data Mining. Analysis of Imprecise Data* (eds.), Kluwer Academic Publishers, 1997.

[6] Manning, C. D. and Schutze, H., *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[7] Pawlak, Z., *Rough sets: Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, 1991.

[8] Polkowski, L. and Skowron, A., *Rough Sets in Knowledge Discovery 2. Applications, Case Studies and Software Systems* (eds.), Physica-Verlag, 1998.

[9] Raghavan, V. V. and Sharma, R.S., "A Framework and a Prototype for Intelligent Organization of Information", *The Canadian Journal of Information Science*, Vol. 11, pp. 88–101, 1986.

[10] Salton, G. and Buckley, C., "Term-Weighting approaches in automatic text retrieval", *Information Processing & Management*, Vol. 4, No. 5, pp. 513–523, 1998.

[11] Skowron, A. and Stepaniuk, J., "Generalized approximation spaces", *The 3rd International Workshop on Rough Sets and Soft Computing*, pp. 156–163, 1994.

[12] Slowinski, R. and Vanderpooten, D., "Similarity Relation as a Basis for Rough Approximations", *Advances in Machine Intelligence and Soft Computing*, P. Wang (ed.), Vol. 4, pp. 17–33, 1997.

[13] Srinivasan, P., "The importance of rough approximations for information retrieval", *International Journal of Man-Machine Studies*, Vol. 34, No. 5, pp. 657–671, 1991.

[14] Willet, P., "Recent trends in hierarchical document clustering: A critical review", *Information Processing and Management*, pp. 577–597, 1988.

Mining Minority Classes From Large Unbalanced Datasets

Head of group	Tu Bao Ho	(Japan Adv. Inst. of Science and Technology)
Collaborator	Duc Duc Nguyen	(Japan Adv. Inst. of Science and Technology)
Collaborator	Saori Kawasaki	(Japan Adv. Inst. of Science and Technology)
Collaborator	Trong Dung Nguyen	(Japan Adv. Inst. of Science and Technology)

1 Introduction

Learning minority classes in unbalanced datasets is a significant problem in both research and applications. It happens that in many domains the user wants to learn only one particular class, and this target class is represented by only a few examples while the others are represented by a large number. Examples of this learning problem are the detection of fraudulent telephone call [4], the detection of oil spills [8] or learning with biological data [11]. Facing with unbalanced datasets, common learning algorithm usually produce unsatisfactory classifiers. It may be caused by the fact that the fundamental assumption of the same class distribution and measures of learning performance built in these learning algorithms are not met or suitable in unbalanced datasets.

Learning minority classes in unbalanced datasets is among issues that were not much previously considered by the machine learning community, is now coming into light [12], [7]. The main approaches to this problem include rebalancing the training data and cost sensitive classification with or without boosting. Naturally, rebalancing techniques include up-sampling of the minority classes [9] or down-sampling of the majority classes [8]. The cost sensitive classification is usually (though not strictly) applied to the divide-and-conquer approach [14], [15]. Methods in this approach typically bias decisions in different directions, as if there were more or fewer cases in a given class, by raising or lowering the cost of a misclassification.

In this paper we introduce a method called LUPC (stands for Learning Unbalanced Positive Class) that can learn minority positive classes in unbalanced datasets with high performance. The main features of LUPC are a combination of separate-and-conquer rule induction [6] with association rule mining [1]. The strength of LUPC lies in its use of the property of unbalanced datasets and adaptive thresholds on accuracy and coverage of rules associated in its different search heuristics. Careful experimental comparative evaluations show that LUPC not only be well-suited in learning a target minority class but also it can learn all classes with high performance.

2 Learning Unbalanced Positive Class

2.1 Preliminaries

Let $A_1, A_2, \ldots A_p$ denote *attributes* with domains $dom(A_1), dom(A_2), \ldots dom(A_p)$ that can be categorical or numeric. The *instance space* X defined by $A_1, A_2, \ldots A_p$ is the Catersian product $dom(A_1) \times dom(A_2) \times \ldots \times dom(A_p)$. An *instance* is denoted by $< x, c(x) >$ where $x \in X$ and c is a function defined over X and has values in a finite set C of classes. In case of supervised data, values of c over X are given *a priori*. Consider the target class in C as the *positive* class C^+ and all other classes as the *negative* class C^-, we then have the training dataset as union of the positive instance subset (denoted by Pos) and the negative instance subset (denoted by Neg). We consider the case when C^+ is minority, i.e., $|Pos| \ll |Neg|$, and the problem is seen as *learning a minority class in a unbalanced dataset*.

A prediction rule for the positive class C^+ is defined as a conjunction of m attribute-value pairs $\bigwedge_{j=1}^{m}(A_{i_j} = v_{i_j}) \rightarrow C^+$ where $i_j \in \{1, 2, \ldots p\}$ and $v_{i_j} \in dom(A_{i_j})$. Denote by $cov(R)$ the set of all training instances covered by a rule R. This set is divided into two subsets of covered instances in Pos and Neg, denoted by $cov^+(R)$ and $cov^-(R)$, that means $cov(R) = cov^+(R) \cup cov^-(R)$. For the purpose of prediction/classification, the task is to find a set of prediction rules for the target positive class, $R^+ = \{R_1^+, R_2^+, \ldots, R_q^+\}$, so that $Pos \subseteq \bigcup_{i=1}^{q} cov(R_i^+)$ and the discovered rules are "best" in terms of high sensitivity as well positive predictive value, and low false positive rate.

When evaluating the classifier performance, there will be some a positive instances correctly classified as positive, but some b positive instances incorrectly classified as negative. On the other hand, some c negative instances will be incorrectly classified as positive and some d negative instances will be correctly classified as negative. These quantities *a, b, c*, and *d* form the *confusion matrix*. Different functions using these four basic quantities have been used to measure performance of classifiers, typically *sensitivity* $= \frac{a}{a+b}$, *specificity* $= \frac{d}{c+d}$, *positive predictive value* $= \frac{a}{a+c}$, *negative predictive value* $= \frac{d}{b+d}$. Performance of learning from supervised datasets is customarily evaluated by *error rate* $= \frac{b+c}{a+b+c+d}$. However, the unbalance of datasets often hinders performance of standard classification methods. When checking the confusion matrix learned by standard methods, we can observe that classifiers induced from unbalanced datasets usually have false negative rate much higher than average error rate, and sensitivity as well positive predictive value are considerably lower than average accuracy.

The accuracy $acc(R)$ of a rule $R \in R^+$ is estimated by $acc(R) = \frac{|cov^+(R)|}{|cov(R)|}$. The error of rule R is defined as $err(R) = 1 - acc(R) = \frac{|cov^-(R)|}{|cov(R)|}$. This accuracy $acc(R)$ and the positive cover ratio $\frac{|cov^+(R)|}{|D|}$ in fact are *antecedent confidence* and *support* of the rule R in the terminology of association rule mining, respectively (Agrawal et al., 1993). Association mining algorithms work with two given thresholds α and

β on accuracy and cover ratio of rules, $0 \leq \alpha, \beta \leq 1$. A rule R is $\alpha\beta$-strong if $acc(R) \geq \alpha$ and $\frac{|cov^+(R)|}{|D|} \geq \beta$.

The following property that we skip the proof will be used to reduce time of scanning the large Neg in generating and selecting candidate rules for C^+. A candidate rule will be eliminated without continuing to scan though large set Neg if at some point this condition holds. Note that $cov+(R)$ can be quickly determined because the size of Pos.

Proposition 1. *Given threshold α, a rule R is not $\alpha\beta$-strong for any arbitrary β if*

$$cov^-(R) \geq \frac{1-\alpha}{\alpha} cov^+(R)$$

2.2 Algorithm LUPC

LUPC is a separate-and-conquer rule induction method that follows the generic separate-and-conquer scheme (Furnkranz, 1999) with improvements to learn minority classes in unbalanced datasets. Firstly, it carries out a search biasing alternatively on accuracy and/or cover ratio with adaptive thresholds. Secondly, it focuses on doing separate-send-conquer induction in the target class with exploitation of the unbalanced property of datasets that allows trying the beam search with a large beam search parameter and one-sided selection. Additionally, it integrates prepruning and postpruning in a way that can avoid overpruning.

2.2.1 Basic ideas and search bias

As the computational quality of rules is measured by two independent measures of accuracy (antecedent confidence) and support, there is not any total order among rules that we can use to guide the search. However, if we fix a threshold for one measure (for example 90% for accuracy) and let the other varies, we can partially order the goodness of rules. An $\alpha\beta$-strong rule R_i is said better than an $\alpha\beta$-strong rule R_j with respect to α if R_i has support higher than that of R_j. Similarly, we define that an $\alpha\beta$-strong rule R_i is better than an $\alpha\beta$-strong rule R_j with respect to β if R_i has accuracy higher than that of R_j

The task of LUPC is not to find all rules satisfying given min_acc and min_cov but to find a subset of "best" rules that cover Pos. The iterative process of finding the best rule among $\alpha\beta$-strong rules and selecting promising non $\alpha\beta$-strong rules depends on the search heuristics specified by the user or by a default strategy. If both min_acc and min_cov are set to high values we can generally find high accuracy and cover ratio rules but they can often cover a small part of Pos. If both min_acc and min_cov are set with low values then discovered rules from the huge set of acceptable rules can cover Pos completely but they are often of high redundancy. LUPC distinguishes three alternatives that occur in practice and that lead to the three corresponding types of search heuristics:

1. *Bias on rule cover ratio.* It is to find sequentially rules with accuracy equal and greater than min_acc but the cover ratio is as large as possible. LUPC sets parameter α as the user specified min_acc and varies adaptive parameter β from highest possible value to the user specified min_cov.

2. *Bias on rule accuracy*: It is to find sequentially rules with cover ratio equal and greater than min_cov but accuracy is as large as possible. LUPC sets parameter β as the user specified min_cov and varies adaptive parameter α from highest possible value to the user specified min_acc.

3. *Alternative bias on rule cover ratio and accuracy.* In this case, LUPC starts with two highest values of α and β and alternatively reduced each of them when fixing the other and applying the corresponding procedure in the above two cases until reaching the stopping condition. This search heuristic is applied when the user do not specify any bias on accuracy and cover ratio.

2.2.2 The Algorithm

The procedure **Learn-positive-rule**

Figure 1 presents the scheme of algorithm LUPC that learns consequently a rule set from Pos and Neg given user-specified minimum accuracy threshold (min_acc) and minimum cover ratio (min_cov).

The procedure **Learn-positive-rule** starts with an empty $RuleSet$ (line 1) and two adaptive parameters α and β on rule accuracy and rule cover ratio, initialized by subroutine **Initialize** (lines 2). This subprogram specifies α or β given the thresholds min_acc, min_cov, and the user specified bias on accuracy or cover ratio. They are set as big as possible (1 for α and the maximum cover ratio of single attribute-value pair available on Pos) and always bigger than or equal to min_acc, min_cov. If the bias is on one accuracy, then β will be set to min_cov, and vise-versa. If the bias in on both accuracy and cover rate or there is no bias specified by the user, then both α or β are set as the biggest value. Lines 3-8 describe a recursive procedure to learn one the best rule among $\alpha\beta$-strong rules, to add it to the $RuleSet$, to remove positive instances covered by this rule under some conditions, and to change adaptively thresholds α and β.

If there are any instances remain in Pos, and α and β are still equal or greater than min_acc and min_cov (line 3), **Learn-positive-rule** calls the subroutine **BestRule** to learn a new rule that is "best" with respect to the user-specified search bias (line 4). If such a rule can be learned successfully (line 5), some positive instances covered by it will be removed from Pos (line 6) and the learned rule is added to the $RuleSet$ (line 7).

The instances removed from Pos, which are covered by the new rule, are those previously covered by $\delta - 1$ learned rules in $RuleSet$. If this rule is unsuccessfully learned, α and/or β will be adaptively reduced by the subroutine **Reduce** (line 8).

Learn-positive-rule$(Pos, Neg, min\_acc, min\_cov)$

1. $RuleSet = \phi$
2. $\alpha, \beta \leftarrow$ **Initialize**(Pos, min_acc, min_cov)
3. while $(Pos \neq \phi$ **and** $(\alpha, \beta) \neq (min\_acc, min\_cov))$
4. $\quad NewRule \leftarrow$ **BestRule**$(Pos, Neg, \alpha, \beta)$
5. $\quad$ if $(NewRule \neq \phi)$
6. $\quad\quad Pos \leftarrow Pos \setminus Cover^{+}(NewRule)$
7. $\quad\quad RuleSet \leftarrow RuleSet \cup NewRule$
8. $\quad$ else **Reduce**(α, β)
9. $RuleSet \leftarrow$ **PostProcess**$(RuleSet)$
10. return$(RuleSet)$

BestRule$(Pos, Neg, \alpha, \beta)$

11. $CandidateRuleSet = \phi$
12. **AttributeValuePairs**$(Pos, Neg, \alpha, \beta)$
13. while **StopCondition**$(Pos, Neg, \alpha, \beta)$
14. $\quad$ **CandidateRules**$(Pos, Neg, \alpha, \beta)$
15. $BestRule \leftarrow$ First $CandidateRule$ in $CandidateRuleSet$
16. return$(BestRule)$

Figure 1: The scheme of algorithm LUPC

The loop between lines 4-8 is repeated until the stopping condition (line 3) holds. The obtained *RuleSet* can be optionally post-processed by **PostProcess**(*RuleSet*) (line 9) before the procedure returns the final *RuleSet* (line 10). The removing of only positive instances covered by new rule (line 6) is an *one-sided selection*. This one-sided selection differs from that in (Kubat and Marvin, 1997) as it removes instances from the minority target class satisfied constraints while leaving the majority class untouched. This strategy is an appropriate adaptation of the separate-and-conquer scheme in learning a minority class as it can keep the accuracy of candidate rules generated from Pos by **CandidateRules** as all information of the negative majority class is kept.

The procedure **Reduce**(α, β) gradually reduces α or β by ratio Δa and Δc that can be changed in applications. The default quantity Δa for reducing α is set as 2%. The default quantity Δc for reducing β is set as 1% of the biggest value of $cov^{+}(A_{ij} = v_{ij})$ for all available attribute-value pairs (A_{ij}, v_{ij}) in Pos.

The procedure **BestRule**

The procedure **BestRule** conducts a search in the rule space to find the "best" in the subset of generated $\alpha\beta$-strong rules given α and β. The **BestRule** is composed of the subroutine **AttributeValuePairs** (line 12) for determining the ordered set $AttributeValuePairSet$ of candidate attribute-value pairs to be used for generating candidate rules, and the subroutine **CandidateRules** (line 14) for determining the ordered set $CandidateRuleSet$ of candidate rules that is set empty at the beginning (line 11). The **CandidateRules** is repeated until **StopCondition** holds and returns the first $CandidateRule$ in the ordered $CandidateRuleSet$ as the $BestRule$.

The **AttributeValuePairs** determines candidate attribute-value pairs from instances remained in Pos as follows. Each attribute-value pair available on Pos is considered as condition part of a rule with a single premise whose conclusion part is C^+. This pair will be considered as a *candidate* attribute-value pair only if it covers more than $\alpha \times \beta \times D$ instances of the actual Pos (a necessary constraint for being an $\alpha\beta$-strong rule). All candidate attribute-value pairs will be then checked on Neg to see if some of them correspond to $\alpha\beta$-strong rules. Lastly, η candidate attribute-value pairs that form $\alpha\beta$-strong rules, ordered by their accuracy or cover ratio depending on the specified search bias, will be selected and added to the ordered set $AttributeValuePairSet$ where η is a parameter specified by the user. If the number of such attribute-value pairs is still less than η, the attribute-value pair corresponding to a non $\alpha\beta$-strong rule but with the highest accuracy or cover ratio will be selected. The process is repeated until obtaining η attribute-value pairs in $AttributeValuePairSet$ or no more attribute-value pairs can be considered.

The subroutine **CandidateRules** generates the ordered set $CandidateRuleSet$ of γ rules from $AttributeValuePairSet$ as follows. The order of rules depends on the specified search bias on their accuracy and/or cover ratio. The set $CandidateRuleSet$ in fact consists of two changeable parts, one contains ordered $\alpha\beta$- strong rules and the other contains ordered non $\alpha\beta$-strong rules. Of course, non $\alpha\beta$-strong rules with cover lower than β will no longer be considered in next steps and so cannot be considered to add to the non $\alpha\beta$-strong rule part. When **StopCondition** does not holds LUPC continues to improve $AttributeValuePairSet$ in the following way: (1) generating new rules by combining each non $\alpha\beta$-strong rules in $CandidateRuleSet$ with attribute-value pairs in $AttributeValuePairSet$, (2) if any of new generated rules becomes $\alpha\beta$- strong it will be inserted to the ordered part of $\alpha\beta$-strong rules in $CandidateRuleSet$. The **StopCondition** holds if one of the following constraints is matched: (1) there is no change in non $\alpha\beta$-strong rules in the previous run of **CandidateRules**, (2) there are already γ $\alpha\beta$-strong rules in $CandidateRuleSet$.

The subroutine **CandidateRules** may require a lot of checks on Neg to see if a generated candidate rule is $\alpha\beta$-strong. Fortunately, thanks to the property in Proposition 1, many candidate rules are quickly rejected if they are found to match the condition $cov^-(R) \geq \frac{1-\alpha}{\alpha} cov^+(R)$ during the scan of Neg. It is easy to count $cov^+(R)$ for each candidate rule R as Pos is small, and we need only to accumulate

Data	Size	*Pos*	Overall Accuracy			Sensitivity			Pos. Pred. Value		
		(%)	See5	See5C	Lupc	See5	See5C	Lupc	See5	See5C	Lupc
ann	898	0.9	0.90	0.92	**0.94**	0.00	**0.50**	**0.50**	0.00	**1.00**	**1.00**
gla	214	4.2	0.69	0.49	**0.79**	0.33	0.00	**1.00**	0.33	0.00	**0.60**
hea	10,000	10.0	**0.90**	0.84	0.86	0.52	0.53	**0.86**	**0.83**	0.39	0.76
hyp	3,613	4.1	0.93	0.97	**0.98**	**0.98**	**0.98**	**0.98**	**0.89**	0.65	0.69
inf	2,380	1.1	**0.95**	0.94	0.86	**1.00**	0.76	**1.00**	**1.00**	**1.00**	**1.00**
sat	6,435	9.7	0.83	0.74	**0.84**	0.57	0.64	**0.68**	**0.71**	0.62	0.62
seg	2.070	14.1	**0.94**	0.93	0.83	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
smo	28,550	5.2	**0.87**	0.67	0.77	0.49	**0.63**	0.45	0.92	0.27	**0.95**
sic	2,800	6.1	**0.98**	0.96	0.97	0.77	**0.84**	0.81	**0.88**	0.71	0.72
fla	1,666	3.6	**0.81**	0.79	**0.81**	0.40	0.13	**0.47**	0.75	**1.00**	**1.00**
edu	10,000	6.1	**1.00**	0.97	0.99	**1.00**	**1.00**	**1.00**	**1.00**	0.95	0.98
sto	6,773	4.5	**0.69**	0.51	0.66	0.01	0.00	**0.12**	0.25	0.00	**0.27**
Avg			**0.88**	0.83	0.87	0.61	0.61	**0.74**	0.73	0.65	**0.79**

Table 1: Performance on the minority target class (with misclassification costs)

the count of $cov^{-}(R)$ when scanning Neg until either we can reject the candidate rule as the constraint holds or we completely go throughout Neg and find the rule has a satisfied accuracy.

Two parameters η and γ can influence on the findings of LUPC. Generally, the higher value η and γ the higher chance to discover better rules.

3 Evaluation

This section reports our experimental comparative evaluation of LUPC. It aims at evaluating the performance of LUPC on (i) learning one target minority class in a unbalanced dataset; and (ii) learning all classes as a classification method.

After inducing a set of rules, LUPC matches testing instances (or unknown instances) in the following way: If a testing instance matched rules of more than one class, LUPC then predicts its class label by the majority vote weighted by the rule confidence. If a testing instance does not match any rule, it will be classified into the default class that is taken as the class of the majority among unclassified training instances.

3.1 Performance in Learning Minority Classes

We carried out an comparative evaluation of performance in learning one target minority class from a unbalanced dataset between LUPC and other methods using cost-sensitive and rebalancing techniques. The aims is to show that LUPC can improve the sensitivity and positive predictive value when learning minority classes and competes with other techniques.

	Sensitivity			Pos. Pred. Value		
Data	See5up	See5down	Lupc	See5up	See5down	Lupc
ann	**1.00**	**1.00**	0.50	**1.00**	0.33	**1.00**
gla	0.33	**1.00**	**1.00**	0.33	**1.00**	0.60
hea	**0.94**	0.88	0.86	0.61	0.52	**0.76**
hyp	**1.00**	**1.00**	0.98	**0.74**	0.46	0.69
inf	**1.00**	**1.00**	**1.00**	**1.00**	0.78	**1.00**
sat	0.61	**0.76**	0.68	0.58	0.53	**0.62**
seg	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
smo	**0.98**	**0.98**	**0.45**	0.55	0.34	**0.95**
sic	0.82	**0.84**	0.81	**0.72**	**0.72**	**0.72**
fla	**0.80**	0.71	0.46	0.48	0.33	**1.00**
edu	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	9.98
sto	0.28	**0.63**	0.12	0.16	0.14	**0.24**
Avg	0.82	**0.87**	0.76	0.69	0.61	**0.79**

Table 2: Performance on minority target class (with up and down sampling)

The experiments were designed as follows. LUPC is compared with See5, the PC commercial descendent of C4.5 [13], and See5 using cost-sensitive and rebalancing (up and down) techniques. The experiments were done on 12 UCI datasets containing minority classes. These datasets include 11 UCI datasets (in parentheses are the dataset abbreviated name and name of the positive class: anneal (ann, 1), glass (gla, tableware), headache (hea, 2), hypothyriod (hyp, hypothyriod), inf (inf, 1), satellite (sat, 4), segmentation (seg, sky), smoking (smo, 1), sick (sic, sick), flare (fla, F), education (edu, 2), and the stomach cancer dataset (sto, dead<90days) collected at the National Cancer Center in Tokyo. The first three columns of Table 1 summaries the name of datasets, the size of datasets, and the percentage of size of the target minority class in each dataset. Each dataset is randomly divided into a fixed stratified training data (2/3) and a testing data (1/3) for the *common* use in See5 and LUPC.

There different ways to assign costs of misclassification. We tried three common ones and chosen the following that gave the best performance of See5 with cost sensitive: the cost of misclassifying members of class C_i to class C_j is set as $\frac{|C_i|}{\sum_k |C_k| - |C_j|}$. For comparing LUPC with rebalancing techniques, we run See5 on each training dataset and its two derived ones by down-sampling and up-sampling. For a fair comparison, See5 and LUPC are run on all datasets with their default parameters: See5 with min CF = 25%, $min\_cover = 2$, and LUPC with $min\_acc = 80\%$, $min\_cover = 2$, the number of candidate attribute value pairs $\eta = 100$ and number of candidate rules $\gamma = 20$. it is easy to try LUPC

Table 1 shows experimental results concerning cost sensitive learning. Columns 4-6 present the overall accuracy on all classes obtained by each method. Columns 7-9 and 10-12 present the sensitivity and the positive predictive value, respectively,

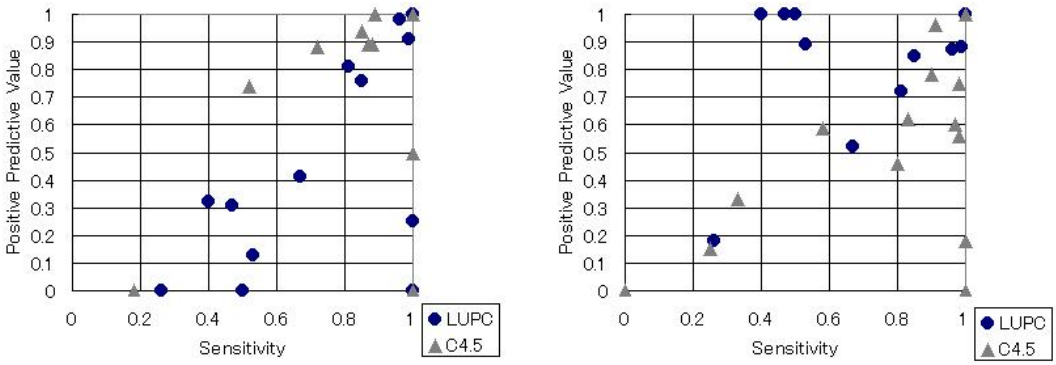


Figure 2: Scatter plot of sensitivity vs. positive predictive value

of each method on the minority positive class.

Table 2 shows experimental results concerning rebalancing techniques. Columns 2-4 present the sensitivity obtained on the minority positive class by LUPC and See5 with up and down sampling. Columns 5-7 present the positive predictive value obtained on the minority positive class by LUPC and See5 with up and down sampling.

Recall that the sensitivity is probability that a positive instance will be classified to the positive class (recall), and the positive predictive value is probability that an instance is positive if it is classified to the positive class (precision). The following conclusions can be drawn:

- Columns 7 and 9 in Table 1 show that the overall accuracy of See5 and LUPC are very comparable (and they are both somehow higher than that of See5 with misclassification costs, column 8), but columns 7, 10, 13 show that the sensitivity and positive predictive value in minority target classes are decreased considerably with See5 (as with some other classification systems in our experiments). It is particularly clear in some datasets such as stomach cancer data (sto).

- Columns 10-12, 13-15 in Table 1 show that See5 with costs improves the sensitivity in minority target classes but it is not clear for positive predictive value. LUPC, however, achieves in many domains sensitivity and positive predictive value higher than those of See5 and See5 with costs.

- Columns 2-4 in Table 2 show that the rebalancing techniques can generally give sensitivity higher than those of LUPC, and down sampling attains somehow sensitivity for minority target class higher up sampling in many domains, but vise-versa for positive predictive value. Columns 5-7 show that LUPC can

Data	Lupc		See5		See5Rules		CBA	
anneal	96.68	± 2.68	91.86	± 0.7	92.78	± 0.8	**97.9**	± 1.32
australian	82.34	± 4.42	84.6	± 1.28	**86.52**	± 1.2	85.94	± 4.19
auto	80.10	± 13.6	**81.20**	± 2.62	79.46	± 2.48	75.60	± 9.4
breastcancer	94.57	± 2.74	95.04	± 0.6	**95.96**	± 0.68	94.4	± 6.23
cleve	**83.77**	± 6.24	78.6	± 2.36	80.08	± 2.24	81.1	± 6.71
crx	84.37	± 4.32	86.8	± 1.08	**86.94**	± 1.16	83.91	± 4.65
diabetes	**78.55**	± 4.41	78.18	± 1.22	76.56	± 5.2	75.61	± 5.51
german	**73.9**	± 3.76	72.58	± 1.02	72.46	± 1.32	73.58	± 4.36
glass	75.74	± 14.7	76.0	± 2.56	74.02	± 3.18	**77.48**	± 6.09
heart	**84.82**	± 2.18	81.26	± 2.74	82.82	± 2.28	82.58	± 10.87
hepatitis	85.3	± 8.19	79.74	± 2.84	83.5	± 2.5	**85.58**	± 7.18
horse	81.4	± 4.37	85.16	± 1.86	**85.48**	± 1.66	82.00	± 3.79
hypothyroid	97.79	± 0.93	**99.2**	± 0.16	99.18	± 0.16	99.05	± 0.61
ionosphere	92.23	± 3.7	89.74	± 1.86	91.5	± 1.46	**93.12**	± 3.88
iris	93.33	± 6.4	**93.44**	± 1.66	94.0	± 1.44	93.25	± 7.12
labor	89.25	± 13.36	83.58	± 5.62	84.14	± 4.76	**94.99**	± 8.07
pima diabetes	**78.5**	± 3.53	77.94	± 1.32	76.8	± 1.22	75.35	± 4.82
sick	97.11	± 1.81	97.92	± 0.26	**97.98**	± 0.26	96.92	± 1.84
sonar	**83.2**	± 7.46	81.12	± 2.43	81.12	± 2.2	80.22	± 9.76
vehicle	70.71	± 3.44	71.6	± 1.32	71.22	± 1.38	**72.67**	± 3.28
waveform	**84.55**	± 1.45	75.48	± 0.52	77.86	± 0.52	81.88	± 0.99
wine	**97.82**	± 3.15	92.66	± 1.94	85.62	± 1.33	**98.33**	± 3.75
Avg	85.73	± 5.31	84.26	± 1.76	84.36	± 1.79	85.52	± 5.20

Table 3: Comparison of accuracy in learning all classes

generally achieve a positive predictive value higher than that of See5 with up or down sampling in the minority target class. LUPC somehow can improve these two measures and balance well these two measures.

- From both Table 1 and Table 2 we can see that LUPC deals well with a minority target class in comparison with cost sensitive learning and rebalancing techniques

Figure 2 gives intuitively views on sensitivity versus positive predictive value of LUPC and See5Cost (right), LUPC and See5 with rebalancing (right) by the scatter plot of these values taken from Table 1 and Table 2, respectively.

3.2 Performance in Learning All Classes

The task of the experiments is to show that although it is developed to learn minority classes in unbalanced datasets LUPC can also achieve performance comparable to the other methods.

We do an experimental comparative evaluation of four systems LUPC, See5, See5Rules and CBA [10])http://www.comp.nsu.edu.sg/∼dm2) on 22 UCI datasets,

some of them are unbalanced as used in the previous experiments, and others are not necessarily unbalanced. The datasets include anneal, Australian, auto, Wisconsin breast cancer (breastcancer), cleve, cxr, diabetes, german, glass, heart disease (heart), hepatitis, horse, ionosphere, iris, labor, pima diabetes, sick, sonar, vehicle, waveform, and wine. The experiments were carried out with 10-fold stratified cross validation. Note that the accuracy of a classification system estimated by stratified cross validation is a random variable, and its value varies with different trials. To have a good estimation of accuracy, we run 10 times of stratified cross validation of See5 and LUPC for each dataset, and take their average as the estimation of accuracy. Though CBA does not make a random partition of a dataset into subsets, and we could not obtain a better estimation of its accuracy on each domain, the objective of our experiments is to show that LUPC can achieve a very comparable performance evaluated on all classes as these well-known classification systems.

Table 3 summarizes the accuracy of four systems on these 22 domains. It can be observed that these systems have a comparable performance. The accuracy of LUPC and CBA are lightly higher than those of See5 and See5Rules but See5 and See5Rules are somehow more stable with smaller standard deviation.

The system LUPC and the datasets used in these experiments can be found and verified at the Web site http://www.jaist.ac.jp/~dungduc/LUPC.

3.3 Discussion

There are two main reasons among others that make LUPC performs well the task of learning positive class in unbalanced datasets.

First is the focus of LUPC on learning positive class and its exploitation of the unbalanced dataset property in different ways: (1) The thresholds on accuracy and cover ratio are determined adaptively to the "hardness" and size ratio of the positive class during the learning process. This contributes to the avoiding of the phenomenon that small classes are blocked by big classes. (2) LUPC selects candidate attribute-value pairs and generates rules from positive instances and uses negative instances to evaluate and select these candidates. The necessary constraint on attribute-value pairs to cover more than $\alpha \times \beta \times D$ instances in Pos can be verified efficiently. The constraint on $\alpha\beta$-strong rules allows LUPC to eliminate considerably candidate rules as $cov^+(R)$ is often small. Recall that it has been experimentally verified that generated rules discovered by beam search with CN2 [2] are often superior to those found by an exhaustive best-first-search. (3) The conditional removal of only positive instances covered by learned rules (one-sided selection) allows LUPC to reduce errors of rules found in further steps as all information on negative class is kept. This cannot be done when learning classes all together. RIPPER [3] learns classes in the order of their prevalence but differs from LUPC as it removes all instances covered by each learned rule.

Second is the effective search of LUPC based on its combination of separate-and-conquer learning and association rule mining. Along the process of separate-

and-conquer induction, LUPC changes adaptively thresholds α and β according to the remaining positive instances so that in each step the best rule is mined in a promising part of the rule space. The chosen heuristics plays an important role in allowing LUPC to vary thresholds on accuracy and cover ratio alternatively from maximum values until the user specified min_accuracy and min_coverage in order to discover a rule as good as possible in each step. This feature is different from other separate-and-conquer induction methods. Moreover, LUPC does not generate all possible rules as CBA in its first phase [10], and this makes LUPC applicable to large datasets.

4 Conclusions

We have introduced the method LUPC developed to learn target minority classes from unbalanced datasets. LUPC is a separate-and-conquer rule induction method using dynamic multiple thresholds on accuracy and cover ratio and the property of unbalanced datasets. Experimental comparative evaluation show that LUPC can learn well a target minority class as well all classes.

References

[1] Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *Inter. Conf. Management of Data SIGMOD'93* (pp. 207–216).

[2] Clark P., & Niblett T. (1989). The CN2 Induction Algorithm. *Machine Learning*, *3*, 261–283.

[3] Cohen, W. (1995). Fast Effective Rule Induction. *Twelfth Inter. Conf. on Machine Learning* (pp. 115–123). San Francisco: Morgan Kaufmann.

[4] Fawcett, T., & Provost, F. (1996). Combining Data Mining and Machine Learning for Effective User Profiling. *Inter. Conf. on Knowledge Discovery and Data Mining KDD'96*, (pp. 8–13). ACM Press.

[5] Frank, E., & Witten, H. I. (1998). Generating Accurate Rule Sets Without Global Optimization. *Fifth Inter. Conf. on Machine Learning ICML* (pp. 144–151). San Francisco: Morgan Kaufmann.

[6] Furnkranz, J. (1999). Separate-and-Conquer Rule Learning. *Journal Artificial Intelligence Review*, *13*, 3–54.

[7] Japkowicz, N. (2000). The Class Imbalance Problems: Significance and Strategies. *AAAI Workshop on Learning in Imbalanced Datasets*, July 2000.

[8] Kubat, M., and Marvin, S. (1997). Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Proc. of the Fourteenth Inter. Conf. on Machine Learning* (pp. 179–186).

[9] Ling, C. X. & Li, C. (1997). Data Mining for Direct Marketing: Problems and Solutions. *Inter. Conf. on Knowledge Discovery and Data Mining KDD-97* (pp. 258–267).

[10] Liu, B., Hsu, W., & Ma, Y. (1999). Integrating Classification and Association Rule Mining. *Fourth Conf. on Knowledge Discovery and Data Mining* (pp. 80–86). New York: ACM.

[11] Muggleton, S. H., Bryant, C. H. & Srinivasan, A. (2000). Measuring Performance When Positive are Rare: Relative Advantage versus Predictive Accuracy – A Biological Case Study. *Proc. of European Conf. on Machine Learning* (pp. 300–312).

[12] Provost, F. (2000). Learning with Imbalanced Data Sets. *AAAI'2000 Workshop on Imbalanced Data Sets.*

[13] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* San Francisco: Morgan Kaufmann.

[14] Turney, P.D. (1995). Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm. *Journal of Artificial Intelligence Research 2* (pp. 369–409).

[15] Ting, K. M. (2000). A Comparative Study of Cost-Sensitive Boosting Algorithms. *Seventeenth Inter. Conf. on Machine Learning* (pp. 983–990).

リポジトリに基づく帰納アプリケーション構築支援環境

研究代表者　山口 高平　(静岡大学情報学部)
研究分担者　和泉 憲明　(静岡大学情報学部)
大崎 美穂　(静岡大学情報学部)
橘 惠昭　(愛媛大学法文学部)
研究協力者　阿部 秀尚　(静岡大学大学院情報学研究科)

はじめに

我々は，帰納アプリケーションの性能に影響を与える手続きバイアスである「学習アルゴリズム」に焦点をあて，帰納学習・マイニングアルゴリズムの選定コストを軽減することを目標とし，帰納学習メソッド (ILMs:Inductive Learning Methods) の体系化としてメソッドリポジトリを構築し，ILMs での操作対象物の体系化としてデータ構造階層を定義した．それらを基に所与のデータセット毎に帰納アプリケーションをメタ学習機構により自動合成するための環境 CAMLET(a Computer Aided Machine Learning Engineering Tool) を開発してきた [9].

今回は，CAMLET に新規のメソッドとそれに伴う新規の制御構造を実装し，さらに仕様探索の効率向上のために仕様実行部分を分散化させた．その上で，ポルト大学 LIACC による StatLog プロジェクト [2] で提供される共通データセットを用いて，CAMLET による正解率の比較と CAMLET のメタ学習による仕様探索について，実験・評価および考察を行う.

帰納学習リポジトリ

帰納アプリケーションを自動的に合成するにはメソッドや操作対象データ構造を適切な粒度で切り出し，体系化する必要がある．CAMLET では 8 種類の代表的な帰納学習システムであるバージョン空間法 [5]・AQ15[4]・ID3[6]・C4.5[7]・ニューラルネットワーク [3]・Bagged C4.5・BoostedC4.5[8]・Classifier System[1] を分析し，メソッドリポジトリおよびデータ構造階層を構築する.

まず，これら 8 種類の帰納学習システムを分析し，制御構造とメソッドリポジトリの上位階層を図 1 のように同定した．ここで，従来のトップレベルのメソッドは 5 種類であったが，“estimate and selecting classifier sets” を加えた，6 種類とした.

メソッドリポジトリ

メソッドリポジトリは図 1 のトップレベルの抽象メソッドを入出力・参照データ構造および機能により，トップレベルのメソッドを具体的なメソッドに階層化したものである．例えば，“estimate and selecting classifier sets” では，はじめに入力される分類器集合の数が単数か複数かによって分岐し，さらにデータセットに対して適用する分類器集合が単数か複数かによって分岐している．メソッドリポジトリは図 2 に示すとおりである.

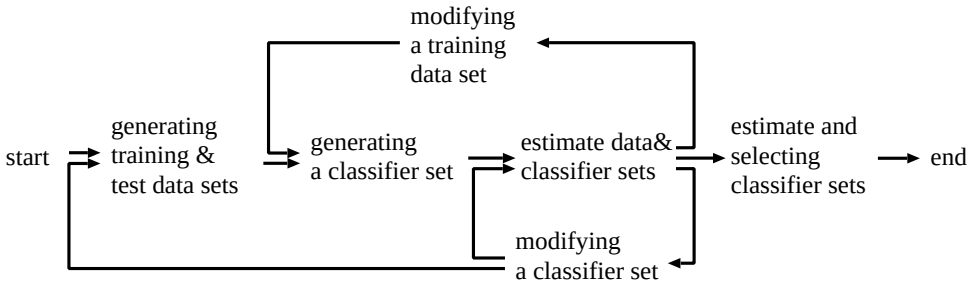


図 1: CAMLET の制御構造テンプレートとトップレベルメソッド

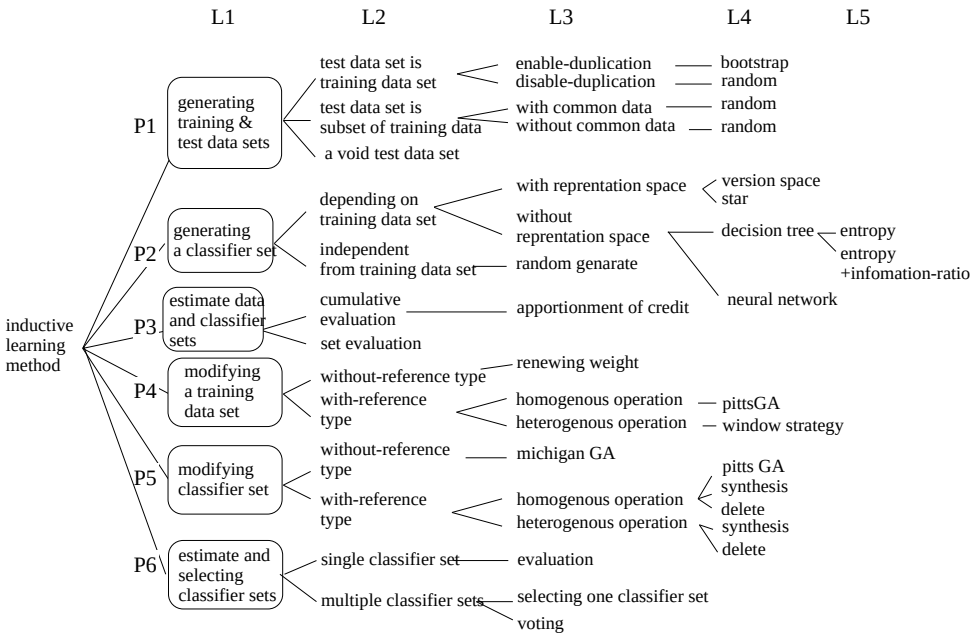


図 2: メソッドリポジトリ階層

データ構造階層

ILMs で操作の対象物 (入力・出力・参照) となるデータ構造を図 3 のように階層化したものを定義する．データ構造の階層はメソッドリポジトリでの階層化に利用される．

CAMLET の基本動作と実行環境

CAMLET はユーザから訓練・テストデータ集合と目標正解率の入力を受け，帰納アプリケーションを生成・テストし，その仕様を正解率とともにユーザに出力として返す．

CAMLET は図 1 の 3ヶ所のフィードバックの有無による 8 種類の制御構造から 1 つを選択し，各トップレベルメソッドをメソッドリポジトリにしたがって，具体的なメソッドを同定し，仕様を決定する．実体化のプロセスでは仕様を実体化し，実行レベルでの実行を依頼する．実行レベルのコンパイル・テストのプロセスで訓練・テストデータ集合を用

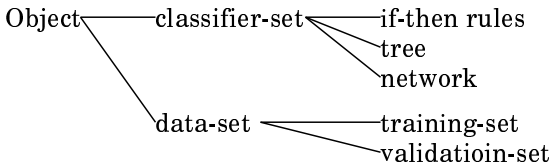


図 3: ILMs におけるデータ構造階層

いて，帰納アプリケーションを仕様に従い組み立て，実行し，正解率を仕様合成レベルに報告する．報告を受けた仕様合成レベルでは正解率が目標正解率に達しているかどうかを調べ，目標に満たない場合は仕様の更新を行う．(図 4)

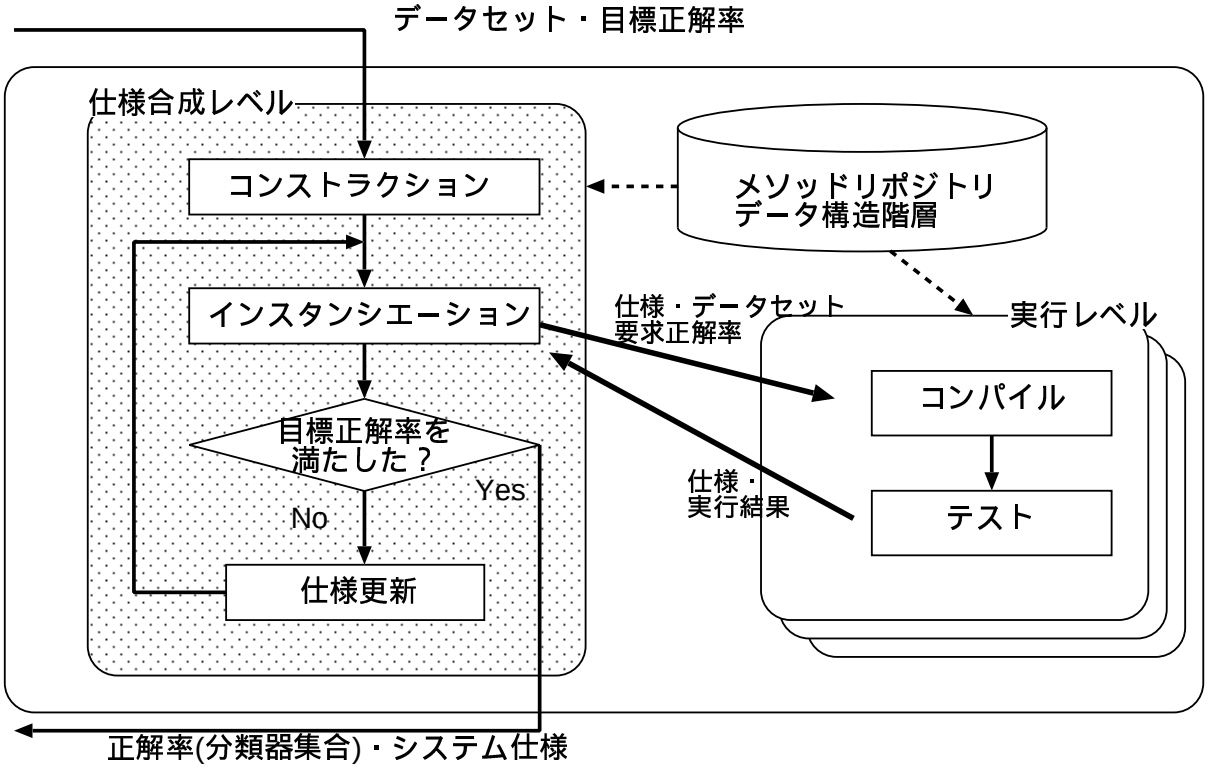


図 4: CAMLET の基本動作

今回新たに設計・実装した分散型 CAMLET は 1 つの仕様合成レベルノードに対し，多数の実行レベルノードを TCP/IP ネットワーク上の計算機に配置する．現在，図 5 に示すアルゴリズムにより合成された仕様を実行している．

CAMLET の評価

CAMLET は UNIX プラットホーム上に Perl で実装され，リポジトリと対応するメソッドは C 言語により実装されている．実験で用いた CAMLET が合成できる仕様は約 2400 種類であるが，制御構造および選択された各メソッドの持つパラメータは固定されている．

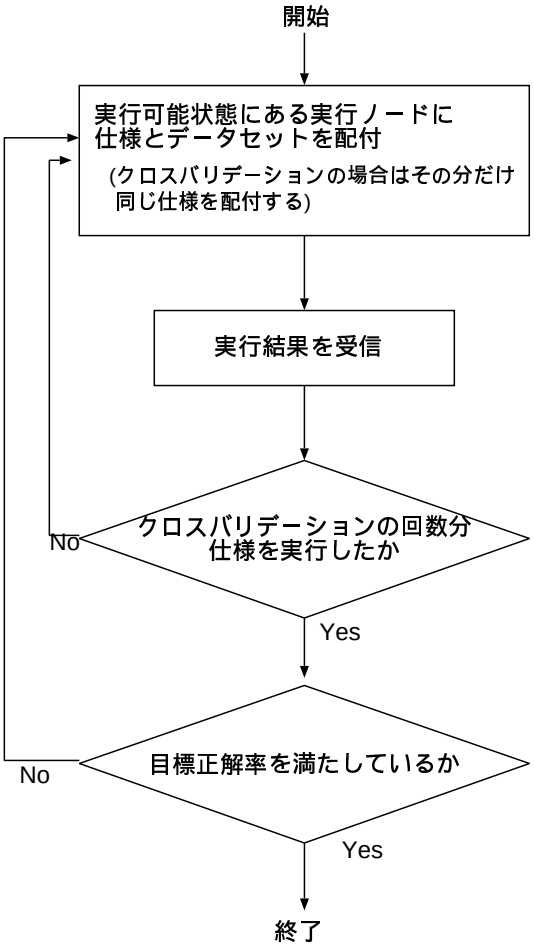


図 5: 分散化を考慮した仕様実行の手順

実験では StatLog プロジェクト[1]により提供される共通データセットを利用する．実験で用いるデータセット 8 種類の詳細は 1 に示すように 4 種類がクロスバリデーションセット，4 種類が訓練・テストデータセットが規定されたデータセットである．クロスバリデーションが指定されている場合，正解率は平均正解率，実行時間は平均実行時間となるが，以下では単に “正解率”，“実行時間” と呼ぶ．

正解率比較実験

正解率比較実験では 8 種類のデータセットと各データセットに対する 24 種類の一般的な学習アルゴリズム・統計システムの正解率と CAMLET により生成される帰納アプリケーションの正解率の比較を行う．比較する正解率はテストデータセットに対する正解率 (= 正しく分類されたデータ数/テストデータ数) である．

CAMLET に与える目標正解率は各データセットでの最高正解率である．仕様生成・更

[1]StatLog プロジェクトでは 24 種類の一般的な学習アルゴリズムと統計システムの 10 種類のデータセットに対する結果が公開されている．CAMLET による実験ではコストマトリクスによる評価が必要な 2 種類のデータセットは用いていない．さらなる情報については http://borba.ncc.up.pt/niaad/statlog/ を参照

表 1: StatLog プロジェクトによるデータセット

データセット	Training	Test
Credit Research for Credit Cards in Australia	10-fold cross validation	
Diabetes of Pima-Indians	12-fold cross validation	
Splice-junction Recognition of DNA Sequence	assigned	assigned
Letter Recognition	assigned	assigned
LANDSAT Satellite Image Recognition	assigned	assigned
Image Segmentation	10-fold cross validation	
Shuttle Control	assigned	assigned
Vehicle Recognition Using Silhouettes	9-fold cross validation	

新はランダムに行い，重複しない 100 個の仕様を試行した時点で 100 個のうち，最も良い正解率となる帰納アプリケーションを比較対象とする．

実験においては，仕様合成機として PentiumIII750MHz × 2 を搭載した WS を 1 台，仕様実行機として UltraSPARCIIi440MHz × 1 を搭載した WS を 54 台使用した．クロスバリデーションを行うデータセットに対してはクロスバリデーションの回数と同じ台数の WS を割り当て，その他のデータセットには複数台の WS を割り当てた．表 2 に今回 CAMLET が合成した帰納アプリケーションによる正解率と StatLog プロジェクトで公開されている結果との比較を示す (C4.5 は平均正解率において 2 番目)．最も良い正解率が得られるまでの時間および探索時間は 1 台で実行したものと仮定し (クロスバリデーションの場合はその回数を同時に実行したと仮定)，該当の仕様にたどり着くまでの時間 (実行時間の合計) としている．図 6〜図 13 は各データセットに対して合成された帰納アプリケーションの仕様である．

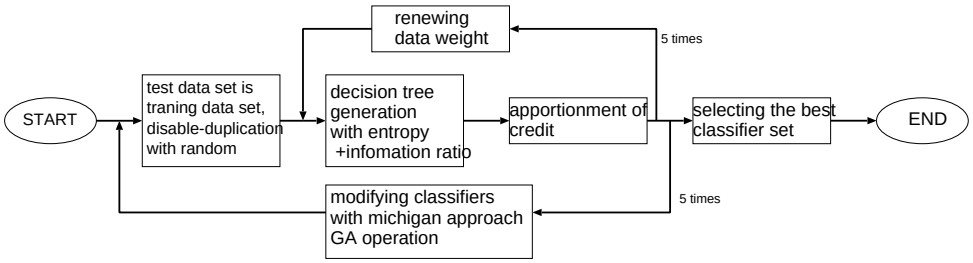


図 6: Credit Research for Credit Card in Australia に対して生成された仕様

CAMLET により合成された帰納アプリケーションの正解率は必ずしも最高正解率を満たしていないが，全体の平均正解率では他の 24 種類のどのアルゴリズムの平均正解率も上回っており，学習システムを固定するよりも優れた結果となる．最終的に選ばれた仕様

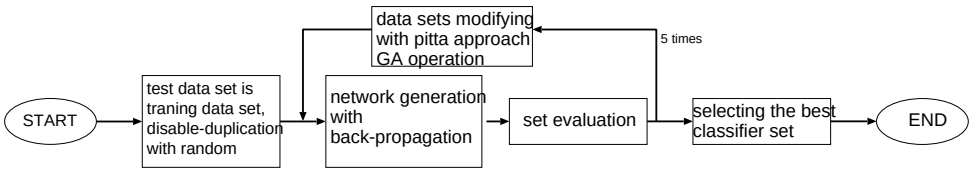


図 7: Diabetes of Pima-Indians に対して生成された仕様

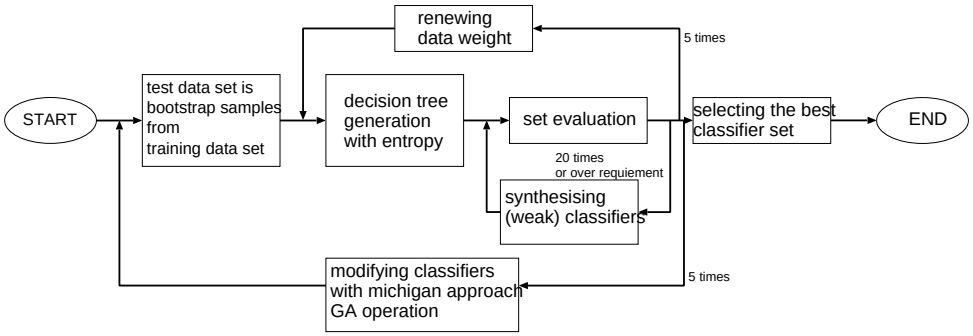


図 8: Splice-junction Recognition DNA Sequence に対して生成された仕様

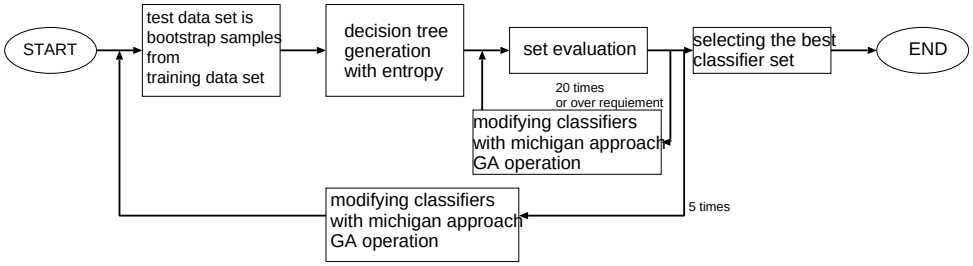


図 9: Letter Recognition に対して生成された仕様

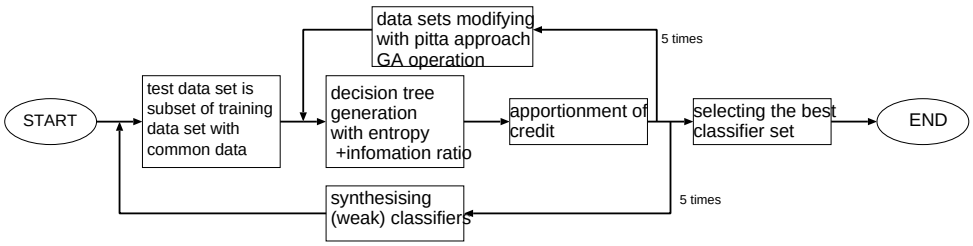


図 10: LANDSAT Satellite Image Recognition に対して生成された仕様

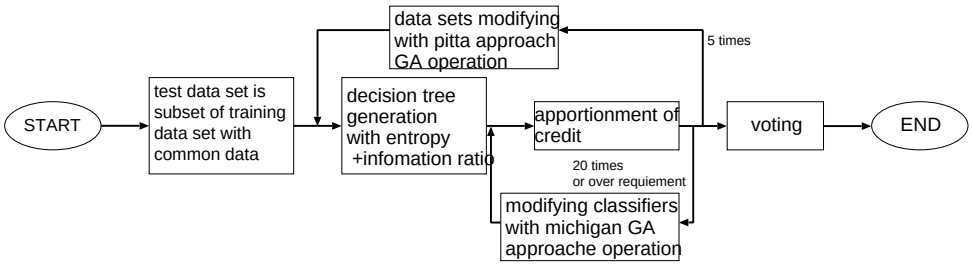


図 11: Image Segmentation に対して生成された仕様

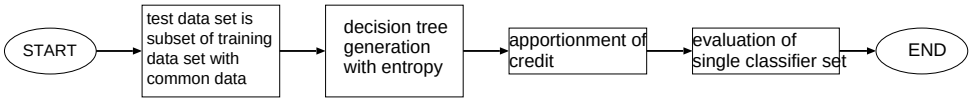


図 12: Shuttle Control に対して生成された仕様

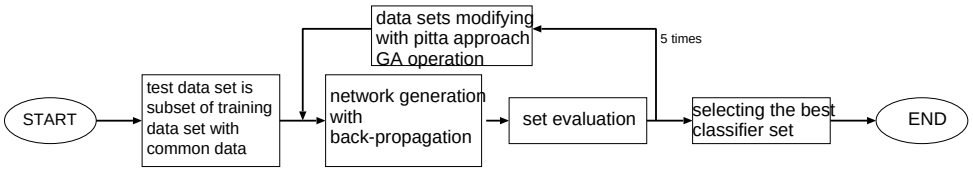


図 13: Vehicle Recognition Using Silhouettes に対して生成された仕様

表 2: CAMLET による結果と StatLog プロジェクトでの結果との比較

Data Set	CAMLET					StatLog		C4.5
	最良仕様			探索				
	正解率 (%)	世代	時間 (sec)	世代	時間 (sec)	正解率 (%)	アルゴリズム	正解率 (%)
Australian	87.3	77	17700	77	17700	86.9	Cal5	84.5
Diabetes	76.4	21	2031	100	20551	77.7	LogDisc	73.0
DNA	95.0	20	19569	100	381689	95.9	Radial	92.4
Letter	82.0	53	151444	100	174106	93.6	Alloc80	86.8
Satimage	88.3	13	214208	100	423704	90.6	KNN	85.0
Segment	96.1	41	71250	100	108608	97.0	Alloc80	96.0
Shuttle	99.8		3893		3893	99.0	NewId	90.0
Vehicle	79.2	23	4034	100	41451	85.0	QuaDisc	73.4
Average	88.0							85.1

を見てみると，8 種類中 7 種類は複数の分類器集合を生成している．その中で最終評価に複数の分類器集合による Voting を用いているのは 1 つだけとなっている．複数の分類器集合を生成し，最終評価では選択された分類器集合の評価を行う仕様では”bootstrap” など，入力された訓練データセットの一部を用いるメソッドが選択されており，データセット全体を使って生成されるモデルより小さなモデルがテストデータセットでの正解率に良い影響を与える事例だと考えられる．また，今回，合成され選択された仕様では Classifier System に由来するランダム分類器生成のメソッドが全く現れていない．これは制御構造・メソッドに与えたパラメータがランダムに生成された分類器集合を十分に訓練できなかったためではないかと考えられる．

メタ学習のための仕様更新ルール学習

上記の実験においては，ランダムに仕様を更新したが，所与のデータセットに対する最良の仕様を効率よく探索するメカニズムが必要となる．そこで，合成仕様と正解率に関するデータを収集し，仕様更新前情報と仕様更新後情報を任意にペアにしたデータセットを構成し，そのデータセットから，仕様更新ルール（メタルール）を学習することを試みた．具体的には，StatLog のあるデータセット (Credit Research for Credit Cards in Australia) に対し，重複しない 1000 個の帰納アプリケーションの仕様をランダムに合成・実行し，仕様と正解率に関する更新前情報と更新後情報のペアから成る訓練データセットを作成し，アプリオリアルゴリズムにより相関ルールを学習させ，未知データセットに適用して，その有用性を検討した．以下，その詳細について述べる．

仕様更新ルール学習のための訓練データセットの作成　仕様更新ルールを学習するための訓練データセットを作成するには，まず，ランダムに合成された 1000 個の仕様から 2 個の仕様を取り出し，それらを更新前仕様 $Spec1$, 更新後仕様 $Spec2$ とする．図 14 に示すように，各仕様はリストで表現され，第 1 要素は制御構造のタイプ番号であり，第 2〜8 要素は制御構造に含まれるメソッド番号が割り当てられる（制御構造は 7 個のメソッドを持

つとしているが，対応する要素にメソッドが存在しない場合は 0 が割り当てられる）．その後，2 つのリストから，対応する要素を結合した新しいリストを作成し，仕様更新ルール学習用の訓練データとする（実装上は，各要素を属性とし，属性－値集合として，データを構成する）．

Spec1: 1,15,24,31,0,0,0,52 → 80.3%

Spec2: 4,13,26,31,0,0,49,54 → 84.6%

データ：1->4,15->13,24->26,31->31,0->0,0->0,0->49,52->54

図 14: 仕様更新ルール学習のための訓練データの構造

ただし，仕様更新ルールは正解率向上を目指しているため，$Spec1$ の正解率よりも $Spec2$ の正解率が高い仕様ペアに限定して訓練データを作成し，今回，1000 個の仕様を合成したが，仕様更新ルール学習用データレコード数は 500,000 以下になっている．また，低い正解率から高い目標正解率に一回で移行できる事は通常まれであり，段階的に向上させる事が必要であると考え，図 15 のように $Spec1$ の正解率によって，仕様更新ルールを適用するためのステージを設定し，訓練データセットを更新前仕様の正解率により各ステージに分割した．さらに，下記の正解率の上昇度によって，クラス値を付与する．以上の作業を実行して，作成された訓練データセットのクラス分布と規模を表 3 に示す．

- ステージの区間の半分以下の上昇 (上昇幅・小)
- ステージの区間内で，ステージ区間の半分以上の上昇 (上昇幅・中)
- それ以上の上昇 (上昇幅・大)

表 3: 訓練データセットの規模

	上昇幅・大	上昇幅・中	上昇幅・小	総インスタンス数
ステージ I		23,537	70,737	94,274
ステージ II	1,530	94,212	145,109	240,852
ステージ III	80,963	21,294	11,035	113,293
ステージ IV	40,919	2,014	2,722	45,656

仕様更新ルールの学習実験　仕様更新ルールを学習するために，今回は，相関ルールを学習するアプリオリアルゴリズムを用いた．最小支持度 (minimum supoort) は，各ステージの総インスタンス数に対する割合として，表 4 のように設定し，各クラス値を結論に持つルールを得るために，各ステージのデータセットに対して，それぞれルールの学習を行

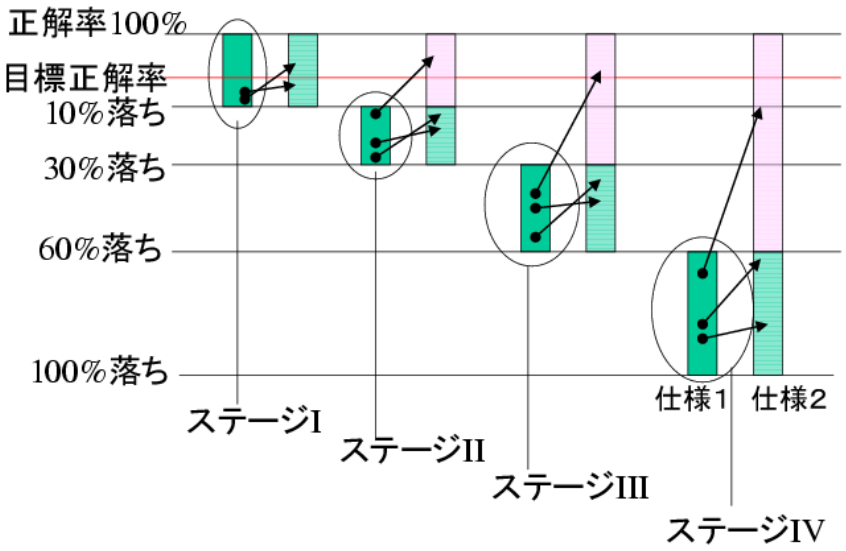


図 15: *Spec*1 の正解率によるステージ設定

う．minimum support をこのように変化させるのは，クラス分布による影響を少なくするためである．minimum support のそれぞれの値は，ステージ毎の” 上昇幅・小” のクラス値を持つインスタンス数を 1 として，その他のクラス値については，各クラス値を持つインスタンス数の比に応じて設定した．なお，” 上昇幅・小” の minimum support は何回かの実験によって調整した値である．最小信頼度（minimum confidence）は，どのデータセットにおいても，75% とした．

表 4: ステージ・上昇幅による minimum support

	上昇幅・大 (%)	上昇幅・中 (%)	上昇幅・小 (%)
ステージ I	7.34	0.33	1.00
ステージ II	0.65	0.65	1.00
ステージ III		1.93	1.00
ステージ IV	15.03	0.73	1.00

この結果得られたルール数を表 5 に示す．

図 16 に，仕様更新ルールの例とその更新過程を図式化したものを示す．点線の四角の部分は，仕様更新ルールが言及する制御構造に含まれない部分であり，”*”の部分は任意の制御構造要素であってよいことを意味する．

表 5: 得られたルール数

	上昇幅・大	上昇幅・中	上昇幅・小
ステージ I	12	0	416
ステージ II	8	0	0
ステージ III			
ステージ IV			

仕様更新ルールの評価　前述の実験から得られた仕様更新ルールを StatLog プロジェクトで提供されている他のデータセットに適用し，評価を試みた．但し，競合が生じた場合は，正解率の上昇幅が最大となるルールを適用する事により，競合解消をはかった．また，目標正解率は，正解率の比較実験と同様，StatLog プロジェクトで提供されている各データセットに記されている最高正解率とした．典型的な結果を図 17,18 に示す (図中，上段はランダム更新の正解率の変化，下段は仕様更新ルール適用時の正解率の変化を示す)．

図 17 においては，仕様更新ルール適用時の方がランダム更新より，合成仕様の最高正解率と最低正解率の乖離が縮小している．しかしながら，仕様更新ルールは常に正解率を向上させている訳でなく，また，50 回仕様を更新させても，目標正解率を上回る仕様を見つける事はできなかった．一方，図 18 では，32 回目の仕様更新時に，目標正解率 (97.0%) に到達できた (図 18 中，★印が 97.1%)．しかしながら，やはり，仕様更新ルールは常に正解率を向上させている訳でなく，約半数の仕様更新が正解率の向上に失敗している．

以上の結果から，現時点では，仕様更新ルールに基づく仕様更新機構は，それほど大きな効果をあげている状況ではない．現在，この問題を解決するために，仕様更新ルールの表現形式を拡張しつつある．すなわち，仕様更新を実行する時に，現在の仕様だけでなく，今までの更新経緯も適用条件とするように拡張中である．

おわりに

本稿では，従来から開発してきた帰納アプリケーション自動構築ツールにコミッティ学習メソッドを追加するとともに，ポルト大学の StatLog プロジェクトから提供されている 8 種類のデータセットを利用して，帰納アプリケーションの自動合成実験を行い，その結果，本ツールで合成された帰納アプリケーションの平均正解率は，StatLog プロジェクトで調査された 24 種類の代表的な帰納アルゴリズムのどの平均正解率よりも高い値を示すことが判った．さらに，効率的な仕様更新を実現するために，相関ルールに基づいて，仕様更新ルールの学習を試み，その結果，ランダム探索よりは安定した仕様更新が実現できることが確認されたが，最良の仕様を効率よく見つけ出す点については不十分であり，仕様更新ルール自身の表現形式について課題が残った．また，本ツールは，帰納アプリケーションに付随するパラメータの探索を行っていないため，パラメータが性能に大きな影響を及ぼすメソッドの改善を得ることができない．今後，各メソッドでのパラメータを同定し，メタ学習による探索を行うこもが課題として残されている．

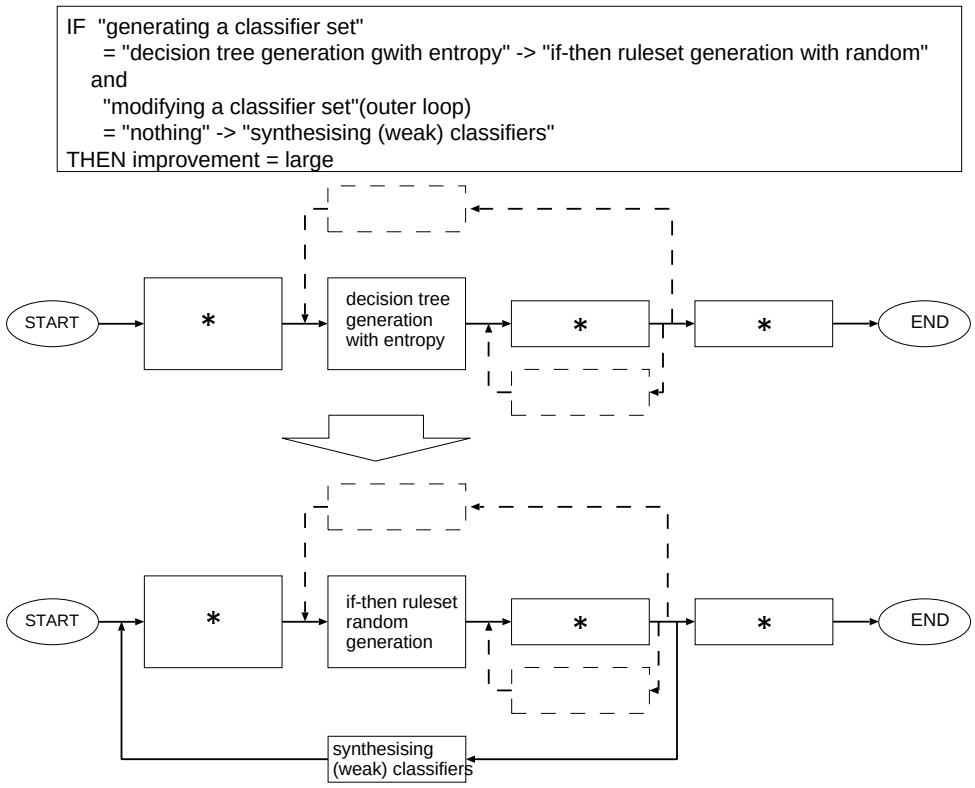


図 16: 仕様更新ルールの例

参考文献

[1] Booker, L. B., Holland, J. H. and Goldberg, D. E., "Classifier Systems and Genetic Algorithms", *Artificail Inteligence*, 40, pp.235–282 (1989).

[2] Brazdil,P. and Henery,R.: "Chapter 10, Analysis of Results", in Machine Learning, Neural and Statistical Classification , D. Michie, D.J.Spiegelhalter and C.C.Taylor (eds.), Ellis Horwood, (1994) pp.175–212

[3] Hinton, G. E., "Learning distributed representations of concepts", *Proceedings of 8th Annual Conference of the Cognitive Science Society*, Amherest, MA. REprinted in R.G.M.Morris (ed.) (1986).

[4] Michalski, R., Mozetic, I., Hong, J. and Lavrac, N., "The AQ15 Inductive Learning System: An Over View and Experiments", *Reports of Machine Learning and Inference Laboratory*, No.MLI-86-6, George Mason University (1986).

[5] "Generalization as Search", *Artificial Intelligence*, 18(2), pp.203-226 (1982).

[6] Quinlan, J. R., "Induction fo Decision Tree", *Machine Learning*, Vol.1, Morgan Kaufmann, pp.81–106 (1986).

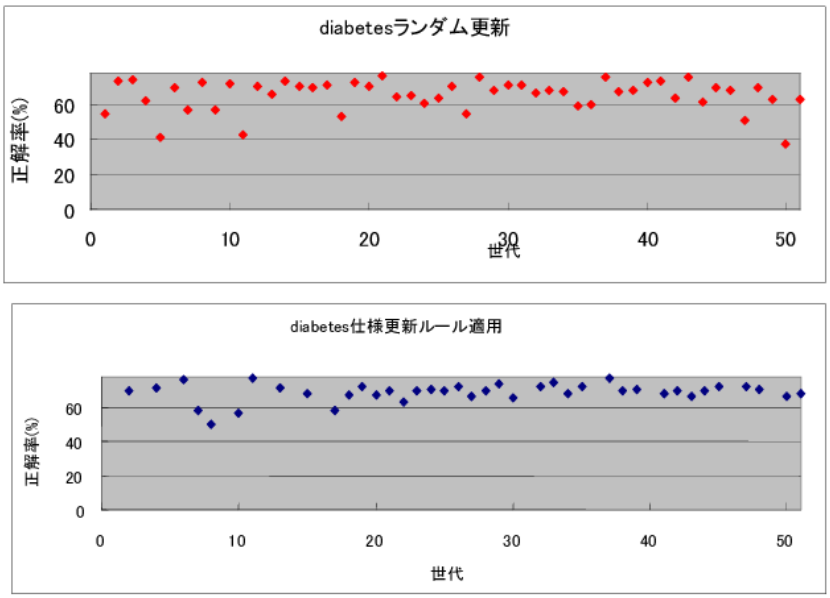


図 17: Diabetes of Pima-Indians におけるランダム更新と仕様更新ルール適用の比較

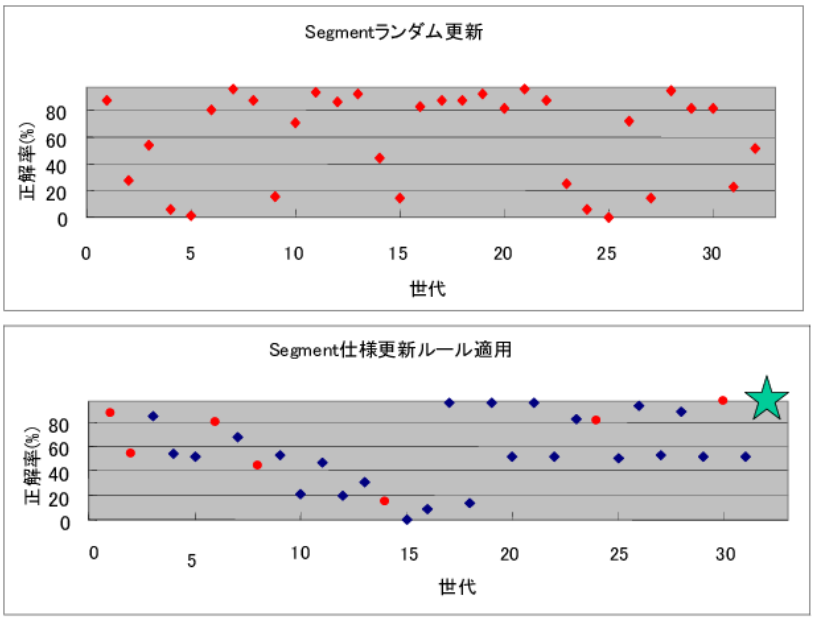


図 18: Image Segmentation におけるランダム探索と仕様更新ルールを適用した探索

[7] Quinlan, J. R., *Programs for Machine Learning*, Morgan Kaufmann (1992).

[8] Quinlan, J. R., "Bagging, Boosting and C4.5", *Proceedings of American Association for Artificial Intelligence* (1996).

[9] 酢山，山口：“オントロジーを利用した帰納アプリケーションの自動合成”，人工知能学会誌 Vol.15, No.1, pp.155–161 (2000)

[10] 阿部 秀尚，山口 高平: 共通データによる帰納アプリケーション自動構築支援環境の評価, 知識ベースシステム研究会，予稿 [SIG-FAI/KBS-J-32], pp. 195-200 (2001).

[11] H.Abe and T.Yamaguchi: Constructing Inductive Applications by Meta-Learning with Method Repositories, to appear in Progresses in DiscoveryScience, State-of-the-Art Surveys, LNCS, Springer-Verlag.

慢性肝炎データセットのクレンジングとマイニングの試み

研究代表者　山口 高平　(静岡大学情報学部情報科学科)
研究協力者　畑澤 寛光　(静岡大学大学院情報学研究科)
佐藤 芳紀　(静岡大学情報学部情報科学科)

背景と目的

医療現場では比較的早い時期から医療情報の電子化が進んでいたため，大学病院をはじめとする大規模な医療施設では入院患者や診察に訪れる患者の検査データが豊富に蓄えられている．数ある KDD の適用分野の中でも医療分野は，(1) 潜在的価値の高い大規模データベースが整備されている，(2) 発見された知見を直ちに臨床にフィードバックさせることが可能である，(3) 公共性が極めて高いため多くの人間がその恩恵を享受できる—ことなどから，今後データマイニングの適用により大きな成果が期待されている分野であると言えよう．

医療現場で構築されたデータベースは，次に述べるような実世界のデータの持つ典型的な特徴を備えている [1]．第一に，患者の検査結果を記録したデータベースは時系列に構築されることが一般的である．時系列データは機械学習による分析が最も困難なデータの一つであり，データの洗浄の他に単一リレーションのデータベースへの変換・データの補充などの前処理が要求される．第二に，医療に限らず自然科学・社会科学においては 1 回の検査で可能な限り多くの情報を収集することが求められるため，一般に極めて多くの属性を持つデータベースが構築される．第三に，検査技術の進歩と相俟って検査項目の追加，あるいは削除が行われるため，検査項目によっては再現性が見られず，現在／過去のデータで発見された知見を検証することができない．また，必然的に多くの欠損値が与えられる．第四に，通常の健康診断で行われているルーチン検査以外の検査項目は医者によるバイアスを受けていることが多い．一般に時間や費用がかかる特殊な検査は病気が進行した段階でしか行われないため，その対象は入院患者や特定の患者に限定される．

本研究ではこれらの情勢を踏まえ，千葉大学病院から提供されたウィルス性慢性肝炎データを対象に，時系列データに対するデータ前処理／知識発見支援機構の開発，及び実験・評価を行った．特に，肝炎の進行具合を示す血液データ（GPT）と検査データとの相関関係の発見を目的とした実験では，肝炎を引き起こすウィルスの周期性について専門医の仮説生成を支援する上で有用な知見を得ることができた [2].

検討内容

本研究が対象とするデータセットはウィルス性慢性肝炎患者の経過追跡データである．データは 1982 年から 2001 年までに肝生検を受けた B 型・C 型の慢性肝炎患者 771 名のプロフィール・検体検査結果情報・肝生検・インタフェロン投与情報などから構成されている．

研究の開始当初，医師からは次のターゲットを提示された:

- 進行の度合いを示す血液データ（GPT）と検査データとの相関関係を発見し，検体

検査データから予後因子[1]を同定する．

- インタフェロン治療例を著効・有効・非有効に分類し，各群を特徴付けるデータを発見する．

本研究では特に前者のターゲット—予後因子の発見—について実験を行った．

以降，提供データの概要，時系列データからの属性構築手法，データの前処理，実験，結果について順番に解説する．

データセットの概要

全部で 5 種類（pt･laboname･labo･bio･ifn）のデータが提供された．このうち，laboname は検体検査結果情報（labo）のメタデータであるため本研究では分析対象から除外した．

分析対象データの概要は以下の通りである:

- 患者基本情報: pt
  患者の性別・生年月日を提供する．レコード数は 771 で 3 種類の属性から構成されている[2]．

- 検体検査結果情報（主に血液・尿）: labo
  患者の検体検査の結果情報を提供する．レコード数は 1,597,146．検体検査結果情報は院内データと外注データに分けられ，最大で 11 種類の属性から構成されている[3]．このうち,「負荷名」「判定結果」「その他」の属性は欠損値が多くマイニングに適さないため,「単位」は知識発見には直接関係しないためそれぞれ削除した．また，院内データの「検査結果」に対して外注データの「検査結果値」を対応付けた．さらに，検査項目のうち，特に「血液型検査」については患者ごとに通常 1 回しか行われず，結果も変化しないことから独立したデータセットとして扱うことにした．

- 肝生検情報: bio
  患者の肝生検[4]の情報を提供する．レコード数は 960 で 8 種類の属性から構成されている[5]．このうち,「検体管理番号」は ID なので削除した．また,「採取施設」も知識 発見に寄与しないため削除した．検査結果についての分類法が数年前から変わったため，新しい分類法である「繊維化」,「活動性」の値がほとんど欠損している．相当数の表記揺れが見られる．

- インタフェロン投与情報: ifn
  インタフェロン[6]を投与した患者の情報を提供する．レコード数は 198 で 4 種類の

[1]治療がうまく行くかどうかを予想するのに利用される要因

[2]「患者 ID」「性別」「生年月日」

[3]「患者 ID」「検査日」「検査項目名」「負荷名」「検査結果値」「単位」「判定内容」「その他」「コメント」「結果評価」「結果項目の子コード」

[4]肝臓の組織を採取して顕微鏡で調べる検査

[5]「患者 ID」「検体管理番号」「肝炎型」「生検年月日」「生検結果」「採取施設」「生検情報（繊維化）」「生検情報（活動性）」

[6]ウィルス性肝炎の特効薬

属性から構成されている[7]．このうち,「投与開始日」は今回提供されたデータではすべて “1 回目” なので削除した．

属性構築手法の選択

予後因子を発見するためのアプローチについて述べる．

提供されたデータセットのうち，主要なマイニング対象となる検体検査結果情報（labo）は患者ごとに数百トランザクションの検査結果が時系列に並べられたものである．検査項目（約千種類）は検査の都度変化し，シーケンスの長さも患者毎にまちまちである．また，期間をおいて検査が再開されるものもあるため，マイニングにはデータの特徴を考慮した適切な前処理が必要となる．

時系列データからルールを帰納する場合，一般的に生データの情報は冗長であるため，時系列データを定量化する属性を新たに構築する必要がある．時系列データのモデル化には様々な手法が存在するが，本研究では時系列変化をクラスタリング・メソッドにより離散化して k 個の名義値で表現する手法を採用する．

ルールを帰納する機械学習スキームは一般化の対象として，目標概念とその概念を記述する属性からなる事例の集合を入力として受け取る．予後因子の同定を目的とする場合，クラスは検査時よりも将来の病理データの傾向を表す名義値として与えられる．同様に属性はクラス設定時より以前の検査データの傾向を記述するものとして与えられる．クラスには病理データの中／長期的トレンドを与えることが最も直感的であり，理にかなっている．属性の構築方法については複数の選択肢が存在する: クラス設定時を基準に相対化したデータを単純に属性として与える方法，時系列変化を特徴ベクトルで記述する方法，時系列変化そのものをパターン化する方法である [3]．このうち，本研究ではパターン化の手法を採用し，属性を構築する．

第一の方法は最も単純であり，予後因子としてクラスの一定期間前のある時点での検査データの値が寄与している場合には有効である．しかしながら，直感的に一時点での検査データの値が病理データのトレンドに寄与しているとは考えにくい．このように複数時点での検査データが同時に寄与している場合には条件が複雑にネストするためルールの帰納や解釈が困難となる．また，一時点での結果に強い相関がみられる場合にも，それがどのような文脈で生じているのかを知る術が無いため専門家による仮説生成の支援には適さない．これらの状況では後述するパターン化の方法が適している．時系列データを抽象化しないことで学習対象を正確に反映したルールが帰納できるが，これは必然的に過適合を招く．また，系列数が多い場合には極端に計算集約的である．

第二の方法は対象とする時系列変化を特徴ベクトルで表現するものである．特徴量としてはデータの水準にしたがって一変量データの要約統計量を使用できる: 離散属性について名義尺度変数，数値属性については順序・間隔・比例尺度変数が対応する．名義尺度とは数値の**区分**のみに注目するものであり，最頻値や度数分布などの変数がある．順序・間隔・比例尺度はそれぞれ数値の**大小関係・差・比**のみに注目するものであり，最大値・最小値・中央値・分散・四分偏差・歪度・尖度などの変数がある．これとは別にパワースペクトル解析により波形の中に含まれる周期変動成分を分離し，それぞれの強さを属性に与

[7]「患者 ID」,「投与回数（N 回目）」,「投与開始日」,「投与終了日」

えることもできる．これらの方法は前者とは異なり，時系列変化の大局的特徴を記述することができるため過適合の問題を回避できる．しかしながら，時系列変化そのものを記述するわけではないためやはり可読性の点に問題がある．また，特徴量の選択は発見的にならざるを得ないため，本質的に次元数の増加をもたらす．次元数を圧縮するためには主成分を抽出するなど副次的な作業を実施しなければならない．

最後の方法は時系列変化から典型的なパターンを抽出してその識別子を名義属性として与える方法である．それぞれの時系列変化はパターンへと変換されることで一定の抽象化が行われる．抽象化の度合いはパターン数で制御できるため，妥当な数を設定することで過適合を回避できる．また特徴ベクトルとは異なり，時系列の表面的な変化を記述するため専門家にとって最も解釈しやすい表現形式であると言える．また，最小で 1 次元で系列を表現できることから多数の系列を扱う際にも計算量の増加による悪影響を受けにくい．しかしながら，パターンの規模と基準値の扱いを検討する必要がある．パターンの規模と基準値を属性として独立させた場合，パターンは純粋に波形の典型的な形状を表す．パターンの規模と基準値はそれぞれ波形の乖離と移動平均値で記述される．直感的にはこれらの 3 属性で波形の表面的特徴を定量化できる．これらの特徴をそれぞれ独立した属性で記述した場合，必然的に次元数が増加する．逆にパターンとして単一の属性で記述した場合，妥当な一般化を行うには非常に多くのパターンが要求される．したがって，これらの特徴の独立の是非を注意深く検討する必要がある．

規模を属性として独立させてパターン化した場合，波形は正規化されるため相対的に見て同じ大きさの変化であれば絶対的な規模の如何に関わらず同等に扱われる．対象となる検査項目を間隔尺度データと仮定すれば，数値の比自体ではなく差のみを重視すればよい．したがって，本研究では波形の規模を属性として独立させずにパターン化を行う．また，基準値を属性として独立させずにパターン化した場合には，絶対零点，すなわち比が定義されるため検査項目の尺度に適合しない．したがって，本研究では基準値として移動平均値を独立させる．

性能は落とし込むパターンの数や欠損値の処理方法に依存している．欠損値の問題は混合分布モデルに基づいた統計的なクラスタリング・メソッドを適用することで回避される．このメソッドでは欠損値はアルゴリズム内部でナイヴベイズ同様に条件付き確率の計算から除外されることで処理されるので，欠損値を平均値・最頻値でフィルタリングしたり線形補完法する手法と比較してロバストであると言える．また，属性の独立性を仮定することで AutoClass[4] と呼ばれる自動的にクラスタ数を決定するための包括的なクラスタリング・スキームが利用できる．このメソッドを適用することで自動的にクラスタ（パターン）の数を同定することが可能である．最終的に切り出される時系列変化の長さをどのように設定するかが問題となる．1 つの方法はメタ学習により決定することである: 交差検定を複数回実行して最も性能の良い長さを選択する．また，J-Measure[5] などの興味深さを定量的に評価する基準を適用して評価値の高いルール群を帰納する長さを選択することも考えられる．

データの洗浄

提供されたデータには相当数の表記揺れが存在しているため以下の方針でデータの洗浄を行う:

- 検体結果情報について
  - 急性肝炎のデータは分析対象外なので削除する.
  - 検査結果値末尾の記号を除去する.
    e.g. $0.982H \rightarrow 0.982$
  - 検査結果値の表記方法を統一する.
    e.g. $20-30 \rightarrow 25$, $10*4.5 \rightarrow 45$
  - 検査結果の大部分が数値で表現されるものについては名義値を欠損値にする.
    e.g. ヨウケツフカ，ニュウリョクミス，キャンセル，ケンサフカ $\rightarrow ?$
  - 検査結果の大部分が名義値で表現されるものについては結果を 2 値化する.
    e.g. $\{$ ヨウセイ，$+$，$etc.\} \rightarrow +1$，$\{$ インセイ，$-$，$etc.\} \rightarrow -1$
- 肝生検情報について
  - 生検結果の名称を大枠で統一する.
    e.g. LC+$Hemangioma\ of\ the\ liver \rightarrow LC$（肝硬変）
  - 急性肝炎のデータを削除する.
  - 活動性の表記を統一する.
- その他データセットについて
  - 「生年月日」「検査日」等の年表記を統一する.
    e.g. *“YYMMDD”* $\rightarrow$ *“YYYYMMDD”*
  - 項目がずれていたり，連結しているレコードを修正する.

特に検体検査結果情報の表記揺れの修正には注意が必要である．本研究では検査結果の時間的変化をパターンとして記述する．名義尺度では二次元のグラフとして描画することが難しいため，検査結果は間隔尺度以上で表現されることが望ましい．したがって，検査の性質上，結果を度数で最頻値で表した方が自然なものも敢えて数値に落した.

検査項目の選択

初期の検体検査結果情報には 957 種類（15,94,390 レコード）の検査項目が存在する．分析対象外の急性肝炎のデータを取り除いても依然 930 種類（1,549,299 レコード）もの検査項目が存在している．本節では入力事例の次元数—検査項目数—を減少させるための方針を述べる.

ほとんどの機械学習アルゴリズムには属性を選択する能力があるが，現実には冗長，あるいは無関係な属性は学習に深刻な影響を及ぼすことが知られている [6]．このため，一

般的にはクラスと相関の強い少数の属性を峻別するために属性選択を絡めた学習が行われる．不適切な属性を削除してデータの次元数を減らすことで学習アルゴリズムの性能は改善するが一般に属性選択は計算集約的である．本研究では出現頻度に基づいて検査項目を剪定した後，属性選択メソッドを併用してルールの発見を行うものとする．

図 1 は検査項目数あたりのレコード数の累計である．全 930 項目のうち上位の約 100 項目で全データの 99% 近くを占め，その後は漸近的に増加している—残りの検査項目はほとんど寄与していない—ことが分かる．最も出現数が多い項目で 46,000 回程度である．10 回に満たないような項目も多数存在している．

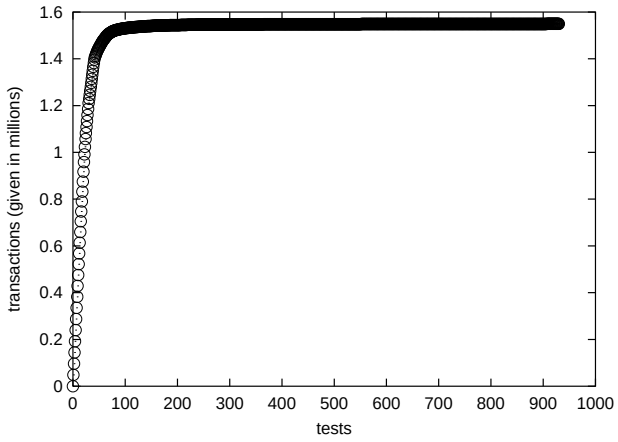


図 1: 検査項目数あたりのレコード数の累計

極端に出現頻度を少ないものを除けば，これらの項目は希少事象を被覆するルールを生成しうるという点で潜在的に有用である．一方で，希少事象は頻度の少なさから母集団を正確に代表しにくく過適合を招きやすいことが容易に想像できる．さらに，検査頻度の少ない検査項目は医者によるバイアスを受けていることが考えられるため，使用の際にはその影響を十分に吟味する必要がある．したがって，理想的には頻度の少ない検査項目のうち，数週間，あるいは数ヶ月に一度実施されているような定期的な検査以外の不定期，あるいは重症患者にのみ実施されているいる検査項目については取り除くべきである．しかしながら，そのような検査項目の選別には多大なコストを要する．また，長期間実施されないことはその観測値が短期的には変化しにくいことを暗示しており，短期的な変化パターンからターゲットを同定する本研究のアプローチに適さない．したがって，本研究では出現頻度のみを基準に検査項目を選り分けるものとし，より緻密なデータの分析に基づいた検査項目の峻別は後続研究に委ねることにする．

実験では簡単に出現頻度が 10000 回未満の検査項目を取り除く．全 930 種類ある検査項目のうち 888 種類，全データの約 1 割 187,870 レコードがこれにあたる．したがって，残りの 42 種類の検査項目を対象にマイニングを実施する．

検査周期の均一化と補完

データの洗浄が終了しても，有効なデータマイニングを行うためにはさらにデータの加工を続ける必要がある．

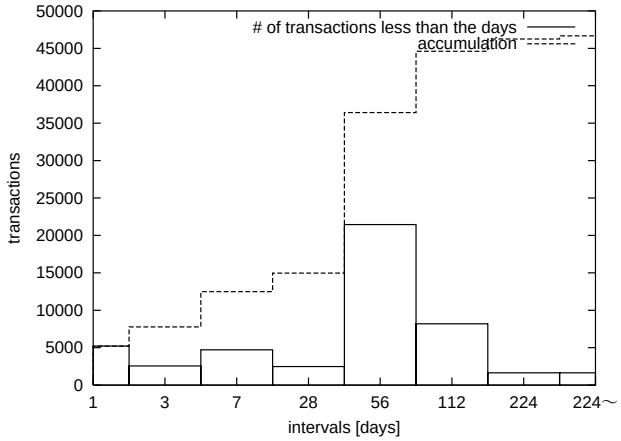


図 2: 検査周期ごとのレコード数

データの表記揺れや偶発的なノイズを取り除き，データセットを検査項目を列に設定して検査日をキーに各検査項目が個別の属性として与えられるようなデータセットに変換する．この状態から属性を構築することも不可能ではない．しかしながら，時系列にソートされた検査記録は不定期であり，系列の異なる記録が一緒くたになっている: 数日おきに検査を受けている患者もいれば，半年，あるいは一年おきに検査を受けているような患者もいる．また，記録は同一の患者のものだけではなく，複数の患者の結果が含まれている．加えて，定期的に検査を受けている患者でも時期によってその期間が短くなったり，あるいは長くなることがある．したがって，妥当な属性を構築するためには検査周期を一定にする必要がある．

データの時系列変化を一般化するためには観測の行われる間隔を一定に保つ必要があるが，問題はその期間の長さ—ウィンドウのプリミティヴの長さ—をどのように設定するかである．一度，区間長を決定するとそれよりも短い周期で行われている検査のほとんどの情報が利用できない．それよりも長い周期で行われている検査についても同様である．この問題に対しては—検査周期を一定に保つ必要がある以上—完全な解決方法は存在しない．したがって，データの分布から判断して適当な長さを設け，場合によっては複数の区間長ごとにデータセットを分割する以外に方法は無い．

図 2 は一定の検査周期を設けてその頻度をヒストグラムで表したものである．これによると 28 日以上 56 日未満に再検査が行われるケースが最も多く，ついで 56 日以上 112 日未満，3 日以上 7 日未満という順になる．本研究では事例数を最大化するような検査周期—ウィンドウのプリミティヴの長さ—を選択する．したがって，28 日（1 ヶ月）が最も多いため，この周期を採用する．周期が 1 ヶ月未満，あるいは 2 ヶ月以上のデータについては線形補完を行う．それぞれ観測が密である空間の値は併合，粗である空間の値は引き延ばすことでデータを補填する．これによりデータセットの情報を最大限利用する．

本研究では以下の手順で検査周期を均一化する．まず，患者 ID 毎に検査日でソートされた検査記録を順に読み出して最初の検査から 1 ヶ月未満に行われた検査をすべてマージする．最初の検査から 1 ヶ月以上離れた検査が現れたところでマージ結果から平均値を求めて検査日をその期間の中心（最初の日付から 14 日後）の日付に設定して検査記録

を出力する．このとき 2 ヶ月以上離れた検査が現れた場合はその期間に応じて空の検査記録を出力する．あらかじめ指定した補完区間数以上離れているか患者 ID が異なる場合は補完を行わず新しい系列として処理する．

検査周期の均一化が終わった段階ではまだ挿入した検査記録は空なので次のパスで補完処理を行う．アルゴリズムは以下の通りである．まず，各検査項目ごとに検査記録を順に読み込んで欠損している箇所，あるいは区間があればその前後の検査記録の値から移動平均値を挿入する．このときあらかじめ指定した補完区間数以上欠損している区間が続く場合は補完を行わない．これは補完する区間が長くなるほどその精度が悪化するためであり，本研究では最長でも 3 区間（3 ヶ月）の補完に留める．

補完方法にはより洗練された方法も存在するが，本研究では分かりやすい見通しを得ることを目的として最も簡単な方法を採用する．図 3 に補完の様子を示す．同一区間の中で複数回検査が行われている場合は結果をマージしてセンタリングを行う．また，検査が行われていない区間では前後の結果の平均を与えることで補完を実施する．

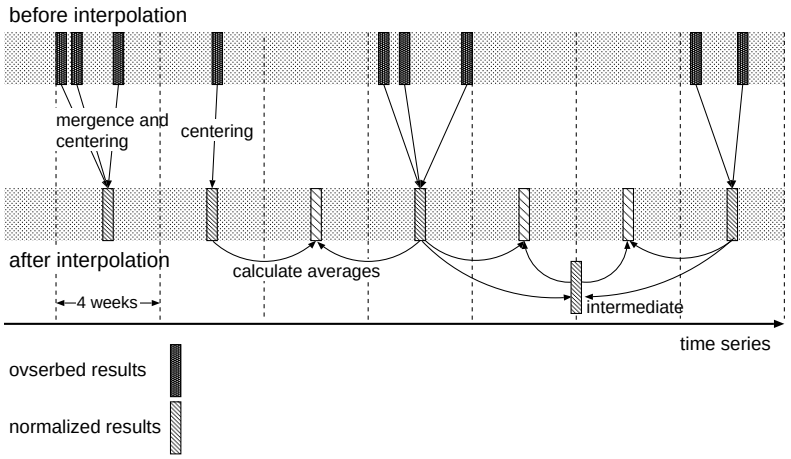


図 3: 検査周期の均一化と検査結果の補完

時系列データの離散化

以下のような手順で時系列データを離散化する．この手順は Das[3] が時系列データからのルール発見に示したフレームワークに則る．

最初に時系列データから切り出すシーケンスの長さを与える．我々はこれをウィンドウ・サイズと呼び，w で表す．シーケンス $s = (x_1, \ldots, x_n)$ が与えられているときウィンドウ・サイズ w で切り出されるサブシーケンスは $s' = (x_i, \ldots, x_{i+w-1})$ となる．シーケンス s の先頭から 1 ずつスライドさせることサブシーケンスの集合 $W(s) = \{s_i | i = 1, \ldots, n-w+1\}$ を求める．

次にクラスタリング・メソッドを用いて各サブシーケンスを離散化する．本研究では EM アルゴリズムを適用するため，ここでは各サブシーケンスが属する分布（母集団）を考える．全サブシーケンスの確率分布（尤度）を最大化するようなパラメータを推定することでクラスタリングを行い，$W(s)$ をクラスタ集合 $C_1, \ldots, C_k$ に変換する．それぞれの

クラスタ C_h に対して記号 a_h を与える．したがって，すべてのサブシーケンスの離散化集合は $D(s) = a_{h(1)}, \ldots, a_{h(n-w+1)}$ となる．a_h はサブシーケンスの時系列変化を表すプリミティヴな形状に対応付けられた名義値であり，目標概念の記述言語に使用される．

時系列データの離散化の手順を図 4 に示す．図 4(a) のグラフは検査結果値（属性値・クラス候補）のシーケンスを表している．サブシーケンスはあらかじめ与えられたウィンドウ・サイズで切り出される．クラスには属性として切り出されたサブシーケンスの直後を起点とするサブシーケンスが与えられる．図 4(b) では (a) で切り出したサブシーケンスをクラスタリングしてパターンに落としている．図 4(c) の表はクラスタリング結果をもとに構築したデータセットを表している．属性値とクラス値にはその系列の典型的な形状に対応した記号が与えられている．

サブシーケンスの正規化

クラスタリングはサブシーケンス間の類似性に基づいて行われる．特にサブシーケンスを多次元のユークリッド空間の点として表現し，その距離に基づいてクラスタリングするメソッドではデータを正規化して各変量の尺度を統一する必要がある．距離関数以外の基準を使用する場合であっても，連続した数値データをクラスタリングする場合にはデータの正規化を実施する必要があるかもしれない．この選択は適用に依存している: パターンが形状のみを表現できればいいのか，あるいはその形の規模や起点の高さ，すなわち基準値も同様に表現しなければならないのかによる．規模の正規化はセンタリング，基準値の正規化はスケーリングと呼ばれ，それぞれサブシーケンスの平均を 0，分散を 1 にすることを意味している．観測値を x_i，平均値を μ，標準偏差 σ で表すと正規化の操作は

$$scale(x_i) = \frac{x_i - \mu}{\sigma}$$

で表される．観測値からサブシーケンス平均値を引くことでセンタリング，標準偏差で割ることでスケーリングが行われる．

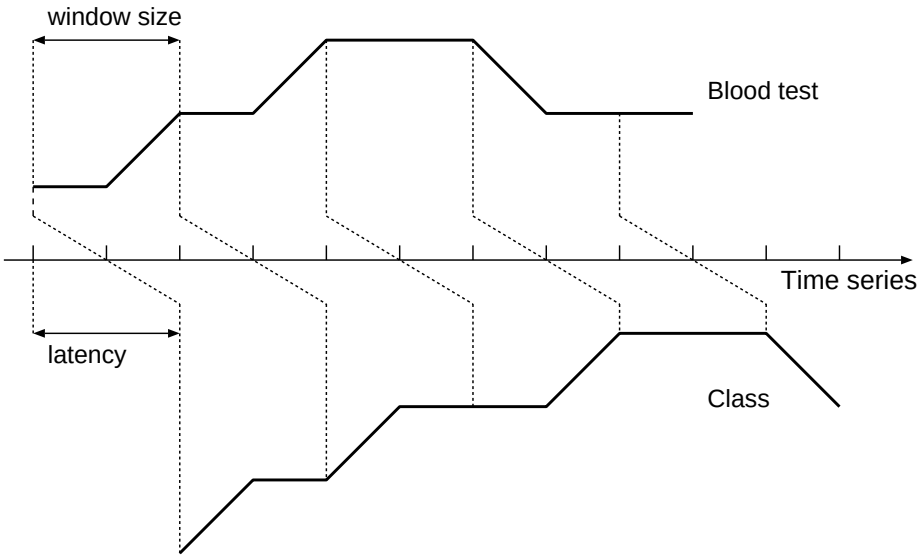
本研究では対象となる時系列データを間隔尺度データとして仮定しているため，数値の差のみが本質的である．したがって，規模の正規化は差の情報が損なわれるため実施しない．また，基準値については正規化しない場合には比が定義されてしまうため正規化を実施する．

クラスタリング・アルゴリズムの選択

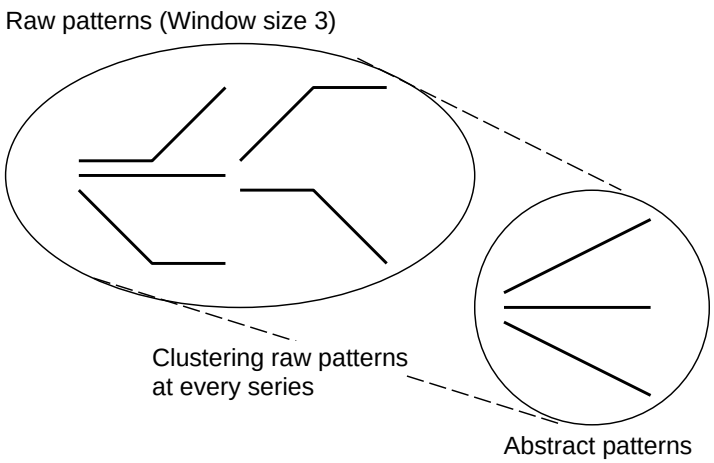
本研究ではシーケンスのクラスタリングに不完全データからの学習アルゴリズムである **EM アルゴリズム**を採用する [7].

k-means に代表されるクラスタリング・アルゴリズムは有限個の根拠から事例を特定のクラスタに決定的に配置するため，単純な問題を除けば，過適合しやすく局所的最適解に陥りやすい．また，類似性の指標にはユークリッド距離の他にも音声認識で使用されている DTW[8]や LCS[9]など様々なものが提案されているが [8]，初期決定に解が左右されるというアルゴリズムの欠点を補うものではない．一方で EM アルゴリズムは大域的最適解

[8]Dynamic Time Warping
[9]Longest Common Subsequence



(a) Forming subsequences by sliding a window
(Using the following subsequence as the Class)



(b) Clustering subsequence according to
the time-seires similarity

Resulting dataset

	B-test	Class
1		
2		
3		
4		
...		

(c) Discretizing time-series by taking
the cluster identifiers correspoinding
to the subsequence

図 4: クラスタリングによる時系列変化の離散化

への収束性に優れていることが経験的に知られており，欠損値やノイズに対しても頑強である．確率的な側面から見た場合，クラスタリングの目標はデータと事前確率が与えられたときに最も尤もらしいクラスタ集合を発見することである．有限個の根拠から完全な決定を下すことは不可能であることから，直感的にも事例を確率的にクラスタに割り当てる EM アルゴリズムは最も自然な方法であると言える．

EM アルゴリズムは正規混合分布モデルに基づいている．このモデルではクラスタは正規分布として表現される．K 個の正規分布 $N(\mu_k, \sigma_k^2), (k = 1, \ldots, K)$ があるとして標本 y はこのうちの一つから与えられるものとする．k 番目の正規分布が選ばれる確率は π_k であるとする．標本が与えられたときにパラメータ π, μ, σ を推定することが問題となる．y の分布は次のように表される:

$$g(y|\pi, \mu, \sigma) = \sum_{k=1}^{K} \pi_k \phi(y|\mu_k, \sigma_k),$$
$$\sum_{k=1}^{K} \pi_k = 1$$

ここで ϕ は正規分布の密度関数

$$\phi(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

である．もし標本が何番目の正規分布から生成されていたかがすべて分かれば問題は自明となる．EM アルゴリズムを適用するために，その番号 z を含めたものを完全データとし，y を不完全データとみなす．完全データ (y, z) の分布は次のように書ける:

$$f(y, z|\pi, \mu, \sigma) = \pi_z \phi(y|\mu_z, \sigma_z)$$

また，独立にこの分布に従う N 個の完全データ
$(y_1, z_1), \ldots, (y_N, z_N)$ が与えられたときの対数尤度は

$$\sum_{i=1}^{N} \log f(y_i, z_i|\pi, \mu, \sigma) = \sum_{i=1}^{N} \log(\pi_{zi} \phi(y_i|\mu_{zi}, \sigma_{zi}))$$

となる．

EM アルゴリズムはパラメータ ξ をある適当な初期値に設定し，**E ステップ**（Expectation step）と **M ステップ**（Maximization step）と呼ばれる二つの手続きを繰り返すことにより ξ の値を逐次更新する方法であり，次のように定式化される:

1. パラメータの初期値を適当な点 $\xi = \xi^{(0)}$ に設定する．
2. $p = 0, 1, 2, \ldots$ に対して次の二つのステップを繰り返す:
   (a) E ステップ: 完全データの対数尤度 $\log f(x|\xi)$ の，データ y とパラメータ $\xi^{(p)}$ に関する条件付き平均を求める．すなわち，
   $$Q(\xi) = E[\log f(x|\xi)|y, \xi^{(p)}]$$
   を計算する．
   (b) M ステップ: $Q(\xi)$ を最大化する ξ を ξ^{p+1} とおく．

肝生検・インタフェロン情報からの属性構築

時系列データを離散化することでクラスと属性が与えられる．しかしながら，その他にも患者のプロフィール・肝生検・インタフェロン情報が利用できるため，最終的にこれらの属性に適切な処理を施して追加することで最終的なデータセットを構築する．

まず，患者基本情報から患者の「性別」「年齢」を追加する.「年齢」についてはクラス設定時を基準にして「生年月日」から算出する．したがって，同じ患者でも切り出したサブシーケンスの時期によって異なる年齢となる可能性がある．

次に，インタフェロン情報からサブシーケンスの時期における患者のインタフェロンの「投与状況」を属性として与える．この属性には 3 つの値 “未投与” ・“投与中” ・“投与済み” を設定する: サブシーケンスの開始時が「投与開始日」以前であれば “未投与”，サブシーケンスの開始時が「投与開始日」と「投与終了日」に挟まれている場合には “投与中”，サブシーケンスの開始時が「投与終了日」以降であれば “投与済み”，インタフェロン投与情報が無い患者の場合には欠損値が与えられる[10]．

最後に，生検結果情報から患者の「生検結果」を追加する．肝生検は血液や尿を調べる検体検査と違って患者一人あたりの検査回数は非常に少なく，多い患者でも 2, 3 回，それも数年の期間をおいている．したがって，時系列データとしてパターン化することはできないのでクラス設定時の近傍（前後 1 年以内）の生検結果の値を属性値として与え，存在しない場合には欠損値とする．

クラス属性には検体検査項目「GPT」のパターンを与え，それぞれ直前の短・中・長期的の検査データのパターンから予測を行う．

結果

検討内容で述べたデータ前処理手法・属性構築手法を実際に提供データに適用し，実験用のデータセットを作成した．

本研究で時系列データから構築した属性はある期間データの特徴を一般化することで生成されたものである．検査項目ごとに同時期の時系列データから切り出した一定の長さのサブシーケンスを，EM アルゴリズムにより同定した混合分布モデルにしたがってその検査項目特有の典型的な変化パターンに落とし込み，対応する名義値に置換したものとその時系列変化の移動平均値をパターンの基準値として属性に与えた．また，パターンの長さとしては 3 種類の長さ—短期（6ヶ月）・中期（12ヶ月）・長期（24ヶ月）—を与えた．したがって，1 種類の検査項目から合計で 6 種類の属性を構築した．また，最大で 42 種類の検査項目が利用可能であるため，データセットの属性数は 240 種類の検査項目属性に肝生検・インタフェロン等の属性を加え，最終的に 250 種類程度となった．

図 5 はクラスタ数を 8 つに設定して同定した GPT の 12ヶ月間の典型的変化パターンである．グラフ右上の凡例にはそのクラスタの事前確率（出現頻度）が%で表現されている．類似した形状を持つパターンが現れているのは，スケーリングを実施していないためであり，y 軸の目盛りを見ることでパターンの規模がそれぞれ異なっていることが分かる．実験では全 42 種類の検査項目のパターンを短期・中期・長期の 3 種類のウィンドウ・サ

[10] “未投与” にしないのは提供された投与情報から漏れている患者が多数存在しているためである

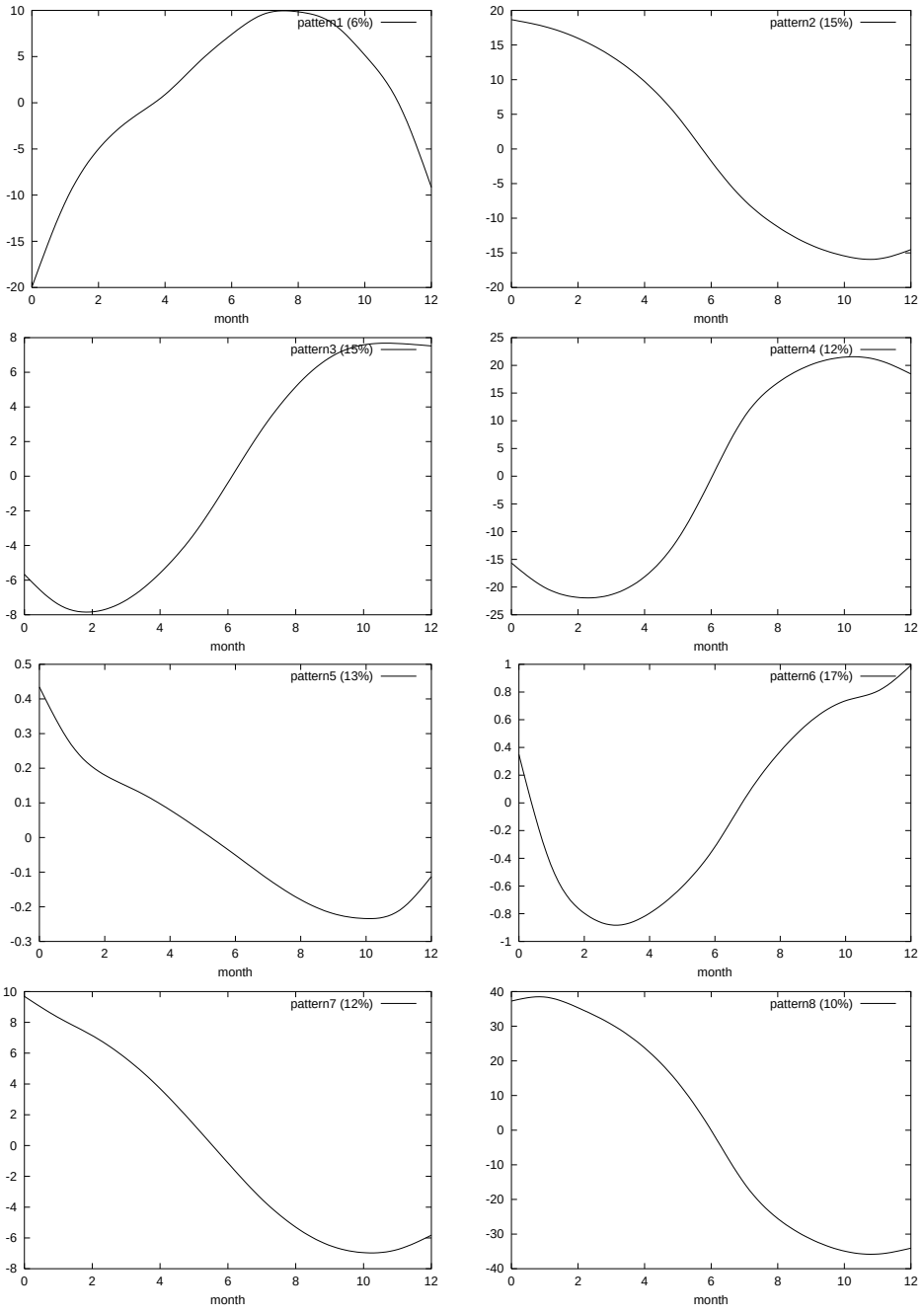


図 5: GPT の 12ヶ月間の典型的変化パターン (対数尤度: -53.9305)

イズで同じように同定し，属性として利用した．

クラスには血液データの中から肝機能を推定する上で有用な指標として利用されている GPT を採用し，各検査項目の直前の変化パターン，及び患者基本情報・肝生検情報等から将来の GPT の変化パターンの同定を試みた．クラスにも属性と同様に短期・中期・長期の 3 種類のウィンドウ・サイズのものを用意した．

以上の手続きにより構築されたデータセットを用いてルール発見を行った．学習スキームには代表的な決定木学習システムである c5.0 を用いた．

図 6-9 のルールは実験により発見されたルールの一例である．これらのルールについて専門医から評価を得ることができた．

学習システムにより発見されたテキストベースのルールはすべてグラフベースのルールへと変換した．クラスタリング時に得られた各クラスタの中心座標の情報，パターンの乖離（最大／最小値の差），基準値（移動平均値）の情報をすべて単一のグラフにプロットして出力するためのスクリプトを実装した．グラフ中央付近の x 軸が 0 となる位置が現在，正の方向が未来，負の方向が過去をそれぞれ表している．ルールに含まれているパターンは波形，基準値は y 軸の中心付近を通る x 軸に平行な直線で表現されている．パターンの乖離の大きさ・基準値の値はそれぞれグラフ右上の凡例に記載されている．尚，パターンについては表示の都合上，分散の大きさが 1 になるようなスケーリングを施してある．

まず，図 6 のルールについては次のようなコメントを頂いた:

> I-BIL（ビリルビン）が高い状態は肝硬変の症状が進んでいることを示すものであり，このルールは直前の 24ヶ月で I-BIL が減少すると GPT が上昇に転じる，という意味に解釈できる．

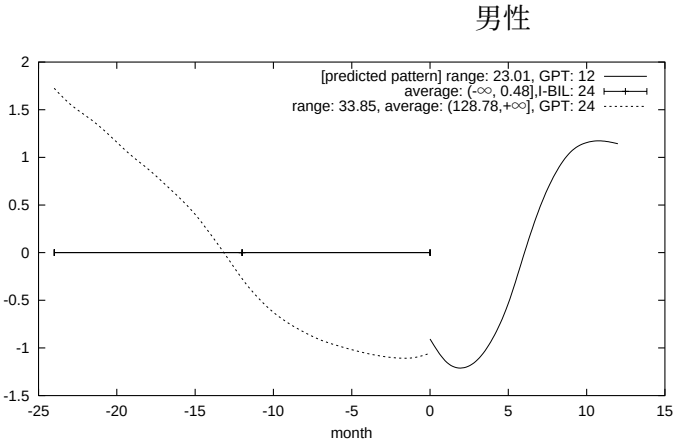


図 6: *precision* : 70.00% (18/25), *recall* : 1.49% (18/1208)

図 7 のルールは「直前 24ヶ月のビリルビンの平均値が高い値を維持し，かつ，TTT が減少すると GPT が減少に転じる」という意味を表す．このルールについては前述のルールと併せて次のようなコメントを頂いた:

医者の感覚では，GPT の値は，ほぼ上昇-下降の周期的な変化を繰り返し，多少の上がり下がりはあるものの，ほぼ一定であると理解されてきた．このルールは GPT の値が上昇から下降，あるいは下降から上昇へと転じる状況を説明するものであり大変興味深い．ウィルスの活動・バクテリアの増殖に周期性があるのか，また，その周期性はウィルスの種類により異なるのか，など，5 〜 10 年の期間で周期性・法則性の示唆ができれば興味深い．

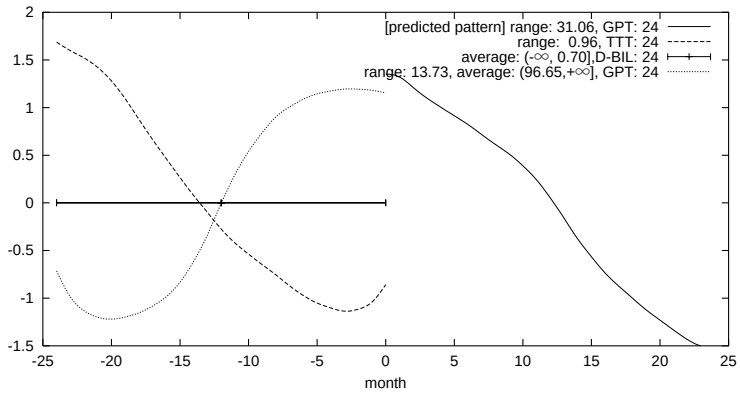


図 7: *precision* : 60.90% (21/34), *recall* : 1.43% (21/1470)

図 8 のルールについては次のようなコメントを頂いた:

乳ビはコレステロール，TG（中性脂肪）が高いことに相当するため，このルールは慢性肝炎がコレステロールとの関連性が高いことを示唆している．また，乳ビは血液の濁り具合を示す医師による主観的な指標であり，ルールに現れてくるのは興味深い．

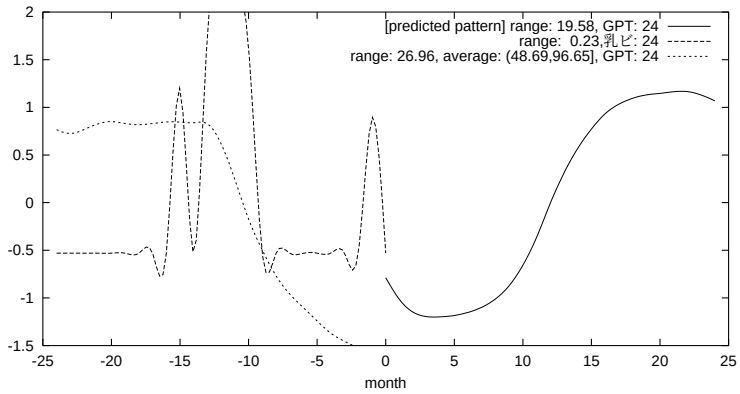


図 8: *precision* : 68.40% (17/24), *recall* : 1.41% (17/1203)

最後に，図 9 のルールについては次のようなコメントを頂いた:

医者の考えと似ているが 6 ヶ月も遅れて GPT が下がるのは意外である．医者の感覚では TTT が下がってから 1, 2 ヶ月後くらいで下がり始めるのが妥当であるが，このようなケースがあってもおかしくはない．もう少し，再現性が高ければ妥当な結果といえるかもしれない．

11 歳以下

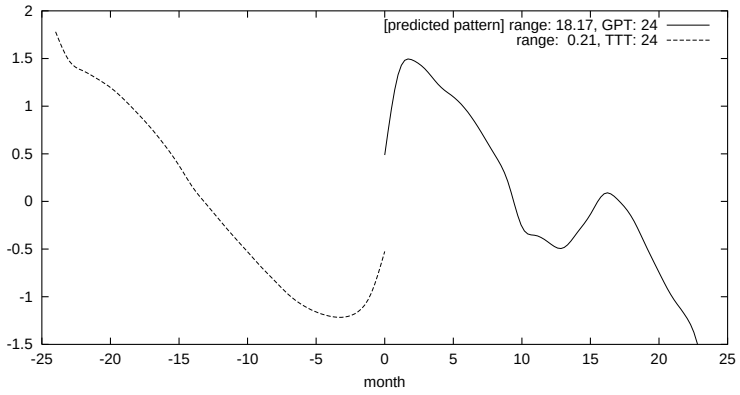


図 9: *precision* : 85.70% (5/5), *recall* : 1.06% (5/470)

時間的制約から，本実験で発見されたルールのうち，実際に医師による評価を得ることができたものは極一部であったが，医学的常識から外れるようなルールは比較的少なく，どちらかといえば医師の経験的知識に近い結果が得られたようである．

また，3 変量以上のパターンが組み合わさったルールについては医師からの評価を得ることができなかった．これは 3 変量以上のパターンの組み合わせが医者の解釈能力を超えてしまうことが原因であったが，この点については今後，ルール発見システムを開発する上で有用なヒューリスティックが得られたという意味で評価できるであろう．

考察

本研究では，ウィルス性慢性肝炎データを対象に，時系列データに対するデータ前処理／知識発見支援機構の開発，及び実験・評価を行った．特に，肝炎の進行具合を示す血液データ（GPT）と検査データとの相関関係の発見を目的とした実験では，肝炎を引き起こすウィルスの周期性について専門医の仮説生成を支援する上で有用な知見を得ることができた．

今後は今回の実験で得られた評価から，肝炎の周期性，及び法則性について有益な示唆をもたらすルールの発見を目指し，手法の改良・実験を行っていく予定である．また，医師から提示されたもう一つのターゲット—インタフェロンを類型化し，各群を特徴付けるデータの発見する—についても研究を開始する予定である．今回の実験では，主にルーチン検査と呼ばれる比較的出現頻度の多い検査項目のみを属性として使用したが，実験結果の評価の過程で医師からルーチン検査以外の検査項目を使用した実験についても強い関心が示された．この点についても，今後の研究で分析を重ねて，カバーしていきたいと考えている．

参考文献

[1] 津本周作: 科学的データベースからの知識発見，チュートリアル JSAI '99, pp.21-38 (1999).

[2] 畑澤 寛光，佐藤 芳紀，山口 高平: 肝炎データセットからの知識発見, 特定 B「情報洪水時代におけるアクティブマイニングの実現」A02・A03 班合同班会議 (2002).

[3] Das, G., et al.: Rule Discovery from Time Series, Proceedings of KDD-98 (1998).

[4] Cheeseman, P., et al.: A bayesian classification system, In Proceedings of the Fifth International Conference on Machine Learning, Morgan Kaufmann (1988).

[5] Smyth, P., et al.: An Information Theoretic Approach to Rule Induction from Databases, IEEE Transactions on Knowledge and Data Engineering, 4 (Aug. 1992), 301-316 (1992).

[6] John, G.H.: Enhancements to the data mining process, PhD Dissertation, Computer Science Department, Stanford University (1997).

[7] 赤穂 昭太郎: EM アルゴリズムの幾何学, 情報処理 Vol.37 No.1, 情報処理学会 (1996).

[8] Gunopulos, D., et al.: Time Series Similarity Measures, KDD 2000 tutorial (2000).

スパイラル的例外性発見に向けて

研究代表者　鈴木 英之進　(横浜国立大学大学院工学研究院)
研究協力者　山田 悠　　(横浜国立大学工学部)

背景と目的

例外と逸脱はデータ集合のごく小さな部分だけに関連するため，機械学習においては従来ノイズとして無視されるか誤認識されてきた．ただしデータマイニングの目的は機械学習の目的より広いと考えられる．例外はしばしば既存知識に疑問を投げかける契機となり，知識を新しい方向に発展させてきたため，知識発見において興味深い対象である．データマイニングにおいては，予測に加えて意志決定の最適化が重要である [13]．例外と逸脱は意志決定の質の向上に貢献するため，発見対象としてより注目されるべきであると思われる．

例外ルールは，一般的なルールに対する形式的な逸脱を表すが，興味深いことが多い．これまで，ユーザの信念に反する例外ルールの効率的な手法が複数個提案されてきた [11, 16, 17]．一方，例外ルールを対応する一般的なルールとペアで発見する手法が，種々の分野で成功を収めて来た [19, 20, 21, 22, 23, 24, 26]．興味深い例外ルールは，相関関係の変化に着目することによっても発見できる [12, 36]．

知識の発見が別の知識の発見につながることは広く知られており，このプロセスが反復される場合も存在する．この現象はスパイラル的発見と名付けられるべきであり，Fayyad らが提案した KDD プロセスの考えに合致する [8]．ただしスパイラル的発見についての知見は少ないためか，このプロセスを自動的データマイニング手法として実現する試みは少ない．このため，この研究課題に関しては予備的研究や事例研究などを積み重ねることが先決である．本稿においては，例外ルールと一般的なルールに関する発見済みのペアを用いて新しいペアを発見する手法について述べ，この手法を知識発見における標準的なデータ集合に適用する．

検討内容

ルールペアの発見

データ集合は n 個の例から構成され，各例は b 個の属性によって記述されるとする．アトムを，名目属性に対する単一値指定か数値属性に対する単一範囲指定と定義する．連言ルールを，前提部が単一アトムかアトムの連言で表され，結論部が単一アトムで表される確率的プロダクションルールと定義する．

著者の一人である鈴木は，ルールペアの集合に関する発見手法を提案した [19, 20, 21, 22, 23, 24]．ここでルールペア $r(x, x', Y_\mu, Z_\nu)$ は，連言ルールのペアとして定義され，それらは一般ルール $Y_\mu \to x$ と例外ルール $Y_\mu \wedge Z_\nu \to x'$ に相当する．

$$r(x, x', Y_\mu, Z_\nu) \equiv \{Y_\mu \to x,\ Y_\mu \wedge Z_\nu \to x'\}$$

ここで，x と x' はそれぞれ名目属性に関する単一アトムであり，属性は同じだが属性値が異なる．各ルールの前提部はアトムの連言 $Y_\mu \equiv y_1 \wedge y_2 \wedge \cdots \wedge y_\mu$, $Z_\nu \equiv z_1 \wedge z_2 \wedge \cdots \wedge z_\nu$ として表される．

本稿の手法は次を満たすルールペアを発見する．

$$\widehat{\Pr}(Y_\mu) \geq \theta_1^{\mathrm{S}},\ \widehat{\Pr}(x|Y_\mu) \geq \theta_1^{\mathrm{F}},\ \widehat{\Pr}(Y_\mu, Z_\nu) \geq \theta_2^{\mathrm{S}},$$
$$\widehat{\Pr}(x'|Y_\mu, Z_\nu) \geq \theta_2^{\mathrm{F}},\ \widehat{\Pr}(x'|Z_\nu) \leq \theta_2^{\mathrm{I}}$$

ただし $\widehat{\Pr}(x)$ はデータ集合における事象 x の割合であり，各 $\theta_1^{\mathrm{S}}, \theta_1^{\mathrm{F}}, \theta_2^{\mathrm{S}}, \theta_2^{\mathrm{F}}, \theta_2^{\mathrm{I}}$ は，ユーザが指定する閾値である．評価指標 $\widehat{\Pr}(Y_\mu), \widehat{\Pr}(x|Y_\mu), \widehat{\Pr}(Y_\mu, Z_\nu), \widehat{\Pr}(x'|Y_\mu, Z_\nu), \widehat{\Pr}(x'|Z_\nu)$ の直観的意味は後に述べる．

初期知識の使用

前節で述べたように，スパイラル的発見は既発見の知識を利用する連続的発見と定義される．利用される知識は提供されるのではなく発見されるため，スパイラル的発見は発見プロセスにおける背景知識の利用とは異なる．知識は利用される前に領域専門家によって評価されるため，スパイラル的発見は探索過程とは異なる．現実のスパイラル的発見をシミュレートすることにより，より有効で効率的な発見手法が開発できると思われる．

既発見の知識を利用する自明な方法として，探索空間を制限することが考えられる．既発見の知識は領域専門家によっていくつかの評価指標に関して順位づけられているため，この方法は可能である．われわれは既に，このような評価指標として妥当性，新規性，意外性，および有用性を用いている [23]．妥当性は発見知識が領域知識に合致する程度を表し，新規性は発見知識が領域知識にとって新しい程度を表す．意外性は発見知識が領域知識にとって部分的には説明できるが常識とは見なされない程度を表す．有用性は発見知識が領域において有用である程度を表す．

われわれは，妥当性と有用性が互いに関連し，新規性と意外性も互いに関連すると考えている．次節で紹介する実験結果はこの考えに合致し，発見知識は４個のグループに分けられる．それらは，全評価指標に高いスコアを示すグループ，妥当性と有用性だけに高いスコアを示すグループ，新規性と意外性だけに高いスコアを示すグループ，および全評価指標に低いスコアを示すグループである．われわれの手法では，発見知識はこの分類法で４グループに分けられ，最初の３グループにおける知識が，探索空間を限定するために用いられる初期知識となる．

最小記述長原理に基づく離散化

実際の発見プロセスでは，発見者は新しい知見を得ようと毎回新しい方法を試みる．探索空間を限定することにより効率性が改善されるため，より計算時間を要する探索手法を用いることが可能となる．このためわれわれは，数値属性の離散化に関してより有効な手法を用いることにした．ルール発見においては，通常数値属性を離散化する必要があり，この過程は発見知識の興味深さに深く関連することが知られている [2]．

Dougherty によれば，離散化手法は教師なし手法と教師つき手法に分類される：前者はクラス属性，すなわちルール発見においては結論部の属性を無視し，後者は考慮する [5]．同様に，離散化手法は大域的手法と局所的手法に分類される：前者は探索前に離散化を行い，後者は探索中に行う．

われわれのこれまでの研究では，時間的効率を考慮して教師なし大域的手法を用いていた [23]．一方今回は，上記の動機に基づき教師つき局所的手法を用いる．具体的には，分類学習において頻繁に用いられる最小記述長原理に基づく手法 [7] を採用した．

発見ルールペア数の削減

再度探索を行うことによりきわめて多数のパターンが生成されるため，発見ルールペア数の削減法を実現した．それらのパターンの多くは，同じ属性の組合せを共有し，数値属性の範囲だけが異なることが，経験的に分かっている．ここで用いた単純な戦略は，各組合せに関して「最良の」ルールペアを得ることである．これを直接的に実現するためには，領域専門家を探索過程に取り込み，発見知識の候補を全て評価してもらえば良い．もっとも，これは明らかに非効率的な方法であるため，前節で紹介した評価指標 $\widehat{\Pr}(Y_\mu)$, $\widehat{\Pr}(x|Y_\mu), \widehat{\Pr}(Y_\mu, Z_\nu), \widehat{\Pr}(x'|Y_\mu, Z_\nu)$, $\widehat{\Pr}(x'|Z_\nu)$ を利用することにした．以下，見やすさのためこれらの評価指標をそれぞれ $s1, c1, s2, c2, c3$ と略記する．

直観的には，$s1$ と $c1$ は一般ルールの一般性と正確性を表す．同様に，$s2$ と $c2$ は例外ルールの一般性と正確性を表す．一方 $c3$ は追加条件 Z_ν が例外ルールの結論部 x' に貢献する度合を表す．よって $c3$ が低いことは例外ルールの意外性が高いことを意味する．

われわれの目的は興味深い例外ルールを求めることであるため，例外ルールは対応する一般ルールよりも重要視されるべきである．ルール発見において，正確性はきわめて重要であると考えられてきた．例えば，ルール発見において先駆的な研究である含意強度 [9, 21] は，ルールの興味深さの指標であり反例の少なさの程度を表す．一方，$c3$ は比較的重要性が低く，その値に関して厳密な程度はあまり要求されず，低ければ良いと考えられる．したがってわれわれは，5 個の評価指標の重要度を，$c2, s2, c1, s1, c3$ の順番であると見なしている．ある特定の属性の組合せについて，これらの評価指標に関して最も重要と見なされるルールペアだけを発見対象とした．例えば，同じ属性の組合せに関して，あるルールペアの $c2$ が最も高ければ，他のルールペアは削除される．もし 2 個のルールペアの $c2$ が最も高ければ，$s2$ の値が高い方だけが発見される．もし 2 個のルールペアの $c2, s2, c1, s1, c3$ に関する値が同じで両方とも最も重要であると見なされれば，両方とも発見される．

結果

初期適用

髄膜炎データの更新版 [25, 34] は，各々が 38 属性で記述される 140 例から構成される．わわれわれの初期適用 [23] においては，ルールペアの各前提部の長さは 1，すなわち $\mu = \nu = 1$ に限定された．他のパラメータは，$\theta_1^{\mathrm{S}} = 0.2$, $\theta_1^{\mathrm{F}} = 0.75$, $\theta_2^{\mathrm{S}} = 5/140$, $\theta_2^{\mathrm{F}} = 0.8$, および $\theta_2^{\mathrm{I}} = 0.4$ に設定された．

領域専門家である津本博士は，妥当性，新規性，意外性，および有用性の観点から発見された各ルールペアを評価した．具体的には，各ルールペアの評価指標値として 1 から 5 までのスコアをつけた．必要と判断した場合，0 をスコアとしてつけた場合もあった．

表 1 に実験結果を示す．表より，この手法は 169 個のルールペアを出力し，妥当性，新規性，意外性，および有用性に関する平均成績はそれぞれ 2.9，2.0，2.0，および 2.7 で

あることが分かる．この実験においては，新規性や意外性が高いルールペアを発見するより，妥当性や有用性が高いルールペアを発見する方が比較的容易であることも分かる．

表 1: 結論部の属性毎にまとめた前手法の平均成績．列 “#” は発見されたルールペアの個数を表す．

属性	#	妥当性	新規性	意外性	有用性
全て	169	2.9	2.0	2.0	2.7
CULT_FIND	4	3.3	4.0	4.0	3.5
CT_FIND	36	3.3	3.0	3.0	3.2
EEG_FOCUS	11	3.0	2.9	2.9	3.3
FOCAL	18	3.1	2.2	2.7	3.0
LOC_DAT	11	2.5	1.8	1.8	2.5
Diag2	72	3.0	1.1	1.1	2.6
KERNIG	4	2.0	3.0	3.0	2.0
SEX	1	2.0	3.0	3.0	2.0
Course (G)	8	1.8	2.0	2.0	1.8
CULTURE	2	1.0	1.0	1.(	
C_COURSE	1	1.0	1.0	1.(	
RISK	1	1.0	1.0	1.(	

前節で述べたように，ここでの発見知識は結論部の属性によって 4 個のグループに分けられる．ここでは 2.5 以上のスコアを高いと見なす．全評価指標に関して高いスコアを示すグループ (CULT_FIND, CT_FIND, EEG_FOCUS)，妥当性と有用性だけに関して高いスコアを示すグループ (FOCAL, LOC_DAT, Diag2)，新規性と意外性だけに関して高いスコアを示すグループ (KERNIG, SEX)，および全評価指標に関して低いスコアを示すグループ (Course (G), CULTURE, C_COURSE, RISK)．より正確に言うと，この傾向はルールペア内の属性に関する組合せに依存する．例えば表 2 に，結論部の属性が CT_FIND であるルールペアに関する傾向を示す．表より，3.5 以上のスコアを高いと見なすと，この場合は 3 個のグループだけがあることが分かる：全評価指標に関して低いスコアを示すグループ (最初の 2 個の組合せ)，妥当性と有用性だけに関して高いスコアを示すグループ (次の 4 個の組合せ)，新規性と意外性だけに関して高いスコアを示すグループ (最後の 2 個の組合せ)．

スパイラル的発見

われわれは前節の発見結果において，評価指標のスコアはルールペアにおける属性の組合せでほとんど説明できることに気づいた．評価指標に関して，数値属性の範囲が重要であることが知られている [2] が，われわれの結果においてそれらは同じであるか，異なる場合には大抵 1 ランクだけである．上記の議論に基づき，重要と見なされる 3 グループから初期知識を選択した．ただし 3.1 より小さいスコアを示すルールペアは無視したた

表 2: 属性の組合せに関する前手法の平均的成績．ただし一般ルールは CT_FIND=normal を予測し，例外ルールは CT_FIND=abnormal を予測する．ルールペアは $r(x, x', y, z)$ で表される

y の属性	z の属性	#	妥当性	新規性	意外性	有用性
(全て)		36	3.3	3.0	3.0	3.2
Cell_Mono	CSF_GLU	2	2.5	3.0	3.0	2.0
Cell_Mono	CSF_CELL	20	3.2	3.0	3.0	3.2
HEADACHE	CSF_PRO	5	3.6	3.0	3.0	4.0
NAUSEA	CSF_PRO	3	4.0	3.0	3.0	4.0
Cell_Poly	Cell_Mono	1	4.0	2.0	2.0	4.0
WBC	CSF_GLU	1	4.0	2.0	2.0	4.0
AGE	FEVER	3	2.0	4.0	4.0	1.0
FEVER	Cell_Poly	1	3.0	4.0	4.0	3.0

め，いくつかのルールペアはスパイラル発見に用いられなかった．これらの重要な 3 グループに関して，平均的スコアと発見ルールペア数を表3 - 5 に載せる．新しく発見されたルールペアは評価のため領域専門家に送付済みである．

発見されたルールペアの例

削減後のルールペア数は 775 であるため，表 5 における 2 番目の組合せに関していくつかの例を示す．まず初期ルールペアを次に示す．ただし “+” は一般ルールの前提部を表す．

```
s1=0.26,c1=0.75,s2=0.043,c2=0.83,c3=0.36
37.6=<BT=<38.8      -> EEG_FOCUS=-
+126=<CSF_PRO=<474 -> EEG_FOCUS=+
```

妥当性：3，新規性：4，意外性：4，有用性：4

このルールペアに基づき，提案手法は次に示す 4 個の新しいルールペアを発見した．これらの各ルールペアは，属性に関する特定の組合せを示す．

```
No.1
s1=0.20,c1=0.75,s2=0.043,c2=0.83,c3=0.40
35.5=<BT=<38.9, 2=<COLD=<9 -> EEG_FOCUS=-
+105=<CSF_PRO=<474          -> EEG_FOCUS=+

No.2
s1=0.26,c1=0.76,s2=0.050,c2=0.86,c3=0.35
```

表 3: 妥当性と有用性に関して高いスコアを示すグループ．ただし ＃ 1，＃ 2，および ＃ 3 はそれぞれ前回発見したルールペア数，前提部にアトムを 1 個追加して今回発見されたルールペア数 (削除前)，今回発見されたルールペア数 (削除後) を表す．評価指標の各スコアは平均値を表す．

y の属性	z の属性	x と x' の属性	＃ 1	妥当性	新規性	意外性	有用性	＃ 2	＃ 3
(全て)			84	3.3	1.8	1.9	3.4	95538	486
Cell_Mono	CSF_CELL	Diag2	43	3.2	1.0	1.0	3.4	0	0
HEADACHE	FEVER	FOCAL	8	3.5	2.1	2.1	3.5	11286	62
HEADACHE	Cell_Mono	FOCAL	3	3.7	2.0	3.0	3.7	24652	150
Cell_Mono	CSF_CELL	CT_FIND	20	3.2	3.0	3.0	3.2	21484	79
HEADACHE	CSF_PRO	CT_FIND	5	3.6	3.0	3.0	4.0	15698	89
NAUSEA	CSF_PRO	CT_FIND	3	4.0	3.0	3.0	4.0	17508	82
Cell_Poly	Cell_Mono	CT_FIND	1	4.0	2.0	2.0	4.0	4508	20
WBC	CSF_GLU	CT_FIND	1	4.0	2.0	2.0	4.0	402	4

表 4: 新規性と意外性に関して高いスコアを示すグループ

y の属性	z の属性	x と x' の属性	＃ 1	妥当性	新規性	意外性	有用性	＃ 2	＃ 3
(全て)			10	2.4	3.8	4.0	2.0	16807	156
Cell_Mono	AGE	Diag2	2	3.0	4.0	4.0	3.0	160	5
AGE	FEVER	FOCAL	1	2.0	4.0	4.0	2.0	1632	21
CRP	WBC	FOCAL	1	2.0	2.0	4.0	1.0	636	30
AGE	FEVER	CT_FIND	3	2.0	4.0	4.0	1.0	913	31
FEVER	Cell_Poly	CT_FIND	1	3.0	4.0	4.0	3.0	5	4
Cell_Poly	AGE	EEG_FOCUS	1	3.0	4.0	4.0	3.0	0	0
HEADACHE	WBC	EEG_FOCUS	1	2.0	4.0	4.0	2.0	385	34
Cell_Poly	CSF_PRO	CULT_FIND	1	2.0	4.0	4.0	2.0	13076	31

表 5: 全評価指標に関して高いスコアを示すグループ

y の属性	z の属性	x と x' の属性	＃1	妥当性	新規性	意外性	有用性	＃2	＃3
(全て)			6	3.4	4.0	4.0	4.0	22770	133
AGE	HEADACHE	EEG_FOCUS	2	3.0	4.0	4.0	4.0	4889	34
BT	CSF_PRO	EEG_FOCUS	1	3.0	4.0	4.0	4.0	171	4
CSF_PRO	FOCAL	CULT_FIND	3	3.7	4.0	4.0	4.0	17710	95

```
35.5=<BT=<39.0, 196=<Cell_Mono=<712 -> EEG_FOCUS=-
+93=<CSF_PRO=<128                   -> EEG_FOCUS=+

No.3
s1=0.34,c1=0.77,s2=0.050,c2=0.86,c3=0.38
35.5=<BT=<39.6, 44=<CSF_GLU=<56 -> EEG_FOCUS=-
+136=<CSF_PRO=<474              -> EEG_FOCUS=+

No.4
s1=0.20,c1=0.75,s2=0.043,c2=0.83,c3=0.36
35.5=<BT=<40.2, 90=<CSF_CELL7=<270 -> EEG_FOCUS=-
+126=<CSF_PRO=<474                 -> EEG_FOCUS=+
```

No. 2 と No. 3 は，評価指標値が改善されていることが分かる．新規の発見プロセスは評価指標値だけに導かれているために，評価指標値が良くなること以外の傾向は観察されなかった．効率性を考えて領域知識の使用と領域専門家の介在を排除したため，この結果はある程度予想されたことである．提案手法の主要な利点は効率性であり，これは初期知識の使用による探索空間の制限と評価指標値に基づく新規発見ルールペア数の削減によって達成されている．

考察

本研究の主な貢献は次の 3 点である．1) 興味深い例外ルールを発見するためのスパイラル発見の形式化；2) 初期知識，最小記述長原理に基づく離散化，および発見ルール数の削減を用いる新手法；および 3) 髄膜炎データ集合を用いた実験的評価．現在，次の 3 点に関する改良に取り組んでいる．1) ユーザの目的などの環境も含むより洗練された形式化，2) より大規模なデータ集合をデータ圧縮 [3, 4, 6, 10, 15] に基づいて効率的に扱う手法，および 3) 医学において新規知識がより求められている肝臓データ集合における評価．

なお紙面の都合上説明は省くが，本課題の基盤技術として有用な機械学習 [14, 28, 33] とデータマイニング [3, 4, 10, 15, 31, 32] に関しても成果をあげた．さらに，本課題に重

要なルール発見の最悪解析にも成果をあげた [29, 30]．特異ルール発見に関する成果については，[37] を参照されたい．

参考文献

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo: "Fast Discovery of Association Rules", *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Menlo Park, Calif., pp. 307–328, 1996.

[2] Y. Aumann and Y. Lindell: "A Statistical Theory for Quantitative Association Rules", *Proceedings of the Fifth ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 261–270, 1999.

[3] 長木悠太，鈴木英之進： 「反復マハラノビスデータ圧縮に基づく高速ブースティング」，第 46 回人工知能学会人工知能基礎論研究会 & 第 54 回人工知能学会知識ベースシステム研究会 合同研究会，pp. 201–206，2001.

[4] 長木悠太，鈴木英之進：「反復データ圧縮型ブースティングの実験的評価」，第 48 回人工知能学会人工知能基礎論研究会，2002 (accepted for publication).

[5] J. Dougherty, R. Kohavi, and M. Sahami: "Supervised and Unsupervised Discretization of Continuous Features", *Proc. Twelfth Int'l Conf. on Machine Learning (ICML)*, pp. 194–202, 1995.

[6] W. DuMouchel, C. Volinsky, T. Johnson, C. Cortes, and D. Pregibon: "Squashing Flat Files Flatter", *Proceedings of the Fifth ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6–15, 1999.

[7] U. M. Fayyad and K. B. Irani: "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning", *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1022–1027, 1993.

[8] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth: "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, pp. 1–34, Menlo Park, Calif., 1996.

[9] R. Gras and A. Lahrer: "L'Implication Statistique: une Nouvelle Methode d'Analyse de Données", *Mathematiques, Informatique et Sciences Humaines*, Vol. 120, pp. 5–31, 1993 (in French).

[10] S. Inatani and E. Suzuki: "Data Squashing for Speeding up Boosting-Based Outlier Detection", *Proc. Thirteenth International Symposium on Methodologies for Intelligent Systems (ISMIS)*, 2002 (accepted for publication).

[11] B. Liu, W. Hsu, L. Mun, and H. Lee: "Finding Interesting Patterns Using User Expectations", *IEEE Trans. Knowledge and Data Eng.*, Vol. 11: 6, pp. 817–832, 1999.

[12] B. Liu, W. Hsu, and Y. Ma: "Pruning and Summarizing the Discovered Associations", *Proc. Fifth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 125–134, 1999.

[13] T. M. Mitchell: "Machine Learning and Data Mining", *CACM*, Vol. 42: 11, pp. 31–36, 1999. (T. M. Mitchell 著，鈴木英之進訳：「機械学習とデータマイニング」，CACM 日本語版，Vol. 1，No. 1，pp. 7–12，2000．)

[14] 長浜光俊，山口直記，鈴木英之進：「粗利と購買履歴に基づく有望顧客の特定」，ビジネスマイニングワークショップ講演論文集，pp. 20–23，2001.

[15] 中本和岐，鈴木英之進:「TWS木を用いた例数圧縮による時系列データの高速クラスタリング」，第48回人工知能学会人工知能基礎論研究会，2002 (accepted for publication).

[16] B. Padmanabhan and A. Tuzhilin: "A Belief-Driven Method for Discovering Unexpected Patterns", *Proc. Fourth Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 94–100, 1998.

[17] A. Silberschatz and A. Tuzhilin: "What Makes Patterns Interesting in Knowledge Discovery Systems", *IEEE Trans. Knowledge and Data Eng.*, Vol. 8: 6, pp. 970–974, 1996.

[18] P. Smyth and R. M. Goodman, "An Information Theoretic Approach to Rule Induction from Databases", *IEEE Trans. Knowledge and Data Eng.*, Vol. 4: 4, pp. 301–316, 1992.

[19] E. Suzuki and M. Shimura: "Exceptional Knowledge Discovery in Databases Based on Information Theory", *Proc. Second Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 275–278, 1996.

[20] E. Suzuki: "Autonomous Discovery of Reliable Exception Rules", *Proc. Third Int'l Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 259–262, 1997.

[21] E. Suzuki and Y. Kodratoff: "Discovery of Surprising Exception Rules Based on Intensity of Implication", *Principles of Data Mining and Knowledge Discovery (PKDD)*, LNAI 1510, Springer, pp. 10–18, 1998.

[22] E. Suzuki: "Scheduled Discovery of Exception Rules", *Discovery Science (DS)*, LNAI 1721, Springer, pp. 184–195, 1999.

[23] E. Suzuki and S. Tsumoto: "Evaluating Hypothesis-Driven Exception-Rule Discovery with Medical Data Sets", *Knowledge Discovery and Data Mining (PAKDD)*, LNAI 1805, Springer, pp. 208–211, 2000.

[24] E. Suzuki: "Mining Bacterial Test Data with Scheduled Discovery of Exception Rules", *Proc. Int'l Workshop of KDD Challenge on Real-World Data (KDD Challenge)*, pp. 34–40, 2000.

[25] E. Suzuki (ed.): *Proc. Int'l Workshop of KDD Challenge on Real-World Data (KDD Challenge)*, 2000 (http://www.slab.dnj.ynu.ac.jp/challenge2000).

[26] E. Suzuki and J. M. Żytkow: "Unified Algorithm for Undirected Discovery of Exception Rules", *Principles of Data Mining and Knowledge Discovery (PKDD)*, LNAI 1910, Springer, pp. 169–180, 2000.

[27] E. Suzuki: "Issues in Organizing a Successful Knowledge Discovery Contest", *Discovery Science (DS)*, LNAI 1967, Springer, pp. 282–284, 2000.

[28] E. Suzuki, M. Gotoh, and Y. Choki: "Bloomy Decision Tree for Multi-Objective Classification", *Principles of Data Mining and Knowledge Discovery (PKDD)*, Lecture Notes in Artificial Intelligence 2168, Springer-Verlag, pp. 436–447, 2001.

[29] E. Suzuki: "Worst-Case Analysis of Rule Discovery", *Discovery Science (DS)*, Lecture Notes in Artificial Intelligence 2226, Springer-Verlag, pp. 365–377, 2001 (erratum: http://www.slab.dnj.ynu.ac.jp/erratumds2001.pdf).

[30] 鈴木英之進：「ルール発見の最悪解析」, 第 46 回人工知能学会人工知能基礎論研究会 & 第 54 回人工知能学会知識ベース システム研究会 合同研究会，pp. 189–194，2001.

[31] 鈴木英之進：「データマイニングにおけるデータ変換」，人工知能学会第 17 回 AI シンポジウム，2002 (accepted for publication).

[32] 鈴木英之進：「データマイニングにおける例外逸脱発見」，統計数理とデータマイニング・発見科学 研究会，2002 (accepted for publication).

[33] 武智文雄，鈴木英之進：「集合属性の利得比上限値に基づく決定木の高速学習」，第 48 回人工知能学会人工知能基礎論研究会，2002 (accepted for publication).

[34] S. Tsumoto et al.: "Comparison of Data Mining Methods using Common Medical Datasets", *ISM Symposium: Data Mining and Knowledge Discovery in Data Science*, pp. 63–72, 1999.

[35] Y. Yamada and E. Suzuki: "Toward Knowledge-Driven Spiral Discovery of Exception Rules", *Proc. 2002 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2002 (accepted for publication).

[36] N. Yugami, Y. Ohta, and S. Okamoto: “Fast Discovery of Interesting Rules”, *Knowledge Discovery and Data Mining (PAKDD)*, LNAI 1805, Springer, pp. 17–28, 2000.

[37] 鍾寧：特異性指向マイニング技法の研究，本号，2002.

特異性指向マイニング技法の研究

研究分担者　鍾寧 (Ning Zhong)　(前橋工科大学工学部)

背景と目的

近年，実世界のデータベースは大規模になり，有用なデータを効率良く利用することが不可能になりつつある．そこで，データベースから知識 (ルール) を発見するデータマイニングに関する研究が重要視されている [1, 3, 4, 5, 6].

データベースから発見されるものは，次の 3 種類に分類することができる．1) 間違った仮説．2) 広く知られていて役に立たない仮説．3) 新しく興味のある仮説．この中の新しく興味のある知識 (仮説) を発見することが，データマイニングの目的である．

今までの手法の中に，統計学的手法を挙げることができる [1, 2, 6]．統計学の分野では，データ集合の中のほかの数値に対して異常であるすべての数値を特異値と呼ぶ．この特異値は，非常に異なっているため，対象間の類似性を決定する際に非特異値よりもずっと強影響を及ぼす．そのためデータ解析を行う前にデータ変換を行い，一定の範囲に収まるようにするか，データ集合から取り除いてしまう．しかし，この取り除かれていた特異データの中にこそ，興味深い知識を発見する手がかりが存在する．そこで，逆に今まで取り除かれていた特異データに注目することで，“新しく興味のある仮説” を発見する可能性が高くなる．

そこで，本研究では特異データを元にしてルールを発見する，特異性指向マイニング技法の開発を行った．また，開発したアルゴリズムを実際にデータベースに適用し，その有効性を確認した [14, 16].

検討内容

特異性指向データマイニングでは，データベース中にある特異データに注目しデータマイニングを行う技法である．ここで言う特異データとは，データベース中に含まれる，ほかとは異なり，かつ相対的に数の少ないデータを指す．特異ルールは，発見された特異データ間の相関性を調べることで発見される．

この特異データを発見する方法の一つとして，本研究で提案した Peculiarity Factor(PF) を利用して発見する方法がある．この PF は次の式で計算することができ，データの特異性が高い (他のデータとは大きく異なり，相対的に数が少ない) 場合は大きな値となり，逆に特異性が低い (他のデータとあまり変わらない) 場合は小さな値となる．

$$PF(x_i) = \sum_{j=1}^{n} N(x_i, x_j)^m \tag{1}$$

ただし，$N(x_i, x_j)$ は属性値間の距離であり，その値は次の様に決定される．

- 連続値の場合，値の差の絶対値を距離とする．
- 記号データの場合，値が記号データの場合，そのままでは距離を決定できないため次のようにして，距離を決定する．

– 背景知識を利用できる場合，背景知識に基づき，それぞれの距離を決定する．
– 背景知識を利用できない場合，異なるデータであれば 1，同じデータであれば 0 と仮定する．

このように決定することで，属性値が連続値であっても，記号データであっても距離を求めることができる．

m は距離の重要度を表すパラメータであり，標準の値は $m = 0.5$ である．$m = 0$ である場合，距離の重要度がもっとも低く，連続値データであっても記号データでありかつ背景知識を利用できない場合と同様の扱いとなる．つまり，値が異なれば 1，同じであれば 0 となる．逆に m が大きくなると (1 を越えると)，距離が小さくても PF はとても大きな値になる．

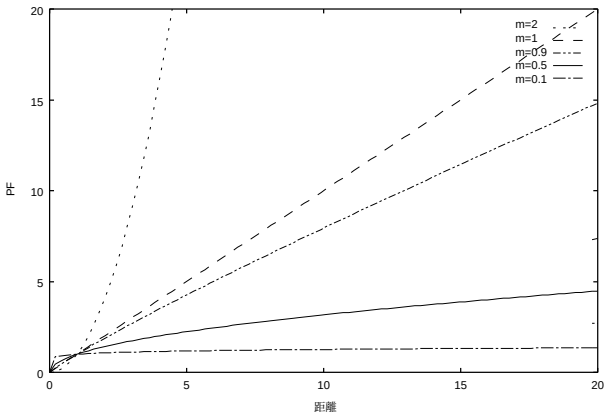


図 1: m を変えたときの距離と PF の関係

データセットに特異データが存在する場合，そのデータの PF の値は大きな値になっている．このことより，PF が平均より一定以上大きなデータが存在する場合，そのデータは特異データであり，そのデータセットに特異データが存在する．このしきい値は，PF の平均値と分散を用いて決定する．

$$\begin{aligned} Threshold &= mean\ of\ PF(x_i) \\ &\quad + \alpha \times standard\ deviation\ of\ PF(x_i) \end{aligned} \tag{2}$$

これにより，コンピュータによる客観的評価で特異データの選択を行うことができる．また，α は人の主観的評価を行うためのパラメータであり，標準では $\alpha = 1$ とする．α を適切に調整することで，人の意志を反映した特異データの選択を行うことができる．

また、データセットに含まれるデータには情報があり、情報を持たないデータはほとんど存在しない。この情報を考慮した上でデータを分類し、特異データを選択する必要がある。PF の値が大きいからといって、それらを一纏めにして特異データとして選択してしまうと、生成した特異ルールの解釈を誤る可能性がある。そのため、あらかじめ元のデータセットからいくつかのクラスタを作成しておき、特異データを発見後、発見された特異データが同じクラスタに属すべきか判別を行う．

このクラスタの作成は、背景知識を利用できる場合は、背景知識を利用して作成する。背景知識を利用できない場合は、連続値であれば最小距離法を用いてクラスタを作成し、記号データであれば異なるデータは異なるクラスタとして扱うようにする．

データベース中のデータや発見された特異データは，必要に応じて情報の粒度の調節を行う．情報の粒度を調節することで，データの抽象化や，概念化を行うことができる．こ

の粒度の調節には，グラニューラコンピューティングを利用する．この技法は背景知識を利用して，情報の粒度を調節する技法である．

グラニューラコンピューティングは，大まかに分けると次の 2 つのグラニューを持っている．1) 基本的なグラニューで，一般的知識として持っているもの．ほとんどのデータセットに適用することができる．2) データセット特有のグラニューで，背景知識として与えるもの．

選択された特異データ間の相関性を調べることで特異ルールを生成する．以上より，特異ルールを生成する手順は次の様になる．

1. クラスタの作成
   データセットに含まれる情報を元にクラスタを作成する．

2. PF の計算
   各属性のデータセットを $X = \{x_1, x_2, \ldots, x_n\}$ と置き，それぞれの PF の値を求める．

3. しきい値の計算
   2 で求めた PF を元にしきい値を計算する．

4. 特異データの選択
   PF がしきい値を越えているデータを特異データとして選択する．

5. データの確認
   特異データの数が十分であれば 7 へ進む．不十分であれば 6 へ進む．

6. データセットの再設定
   データセットから先ほど選択した特異データを抜き出し，新たなデータセットを作成する．その後，2 へ戻る．

7. 情報の粒度の調節
   1 で作成したクラスタをもとに，特異データを修正する．また，グラニューラコンピューティングを用いて，情報の粒度の調節を行う．

8. 特異ルールの生成
   今までに得られた特異データから特異ルールを生成する．

また，本研究では、特異ルールを相関ルールや例外ルールと形式的に比較・分析し，特異性指向マイニングの理論的根拠を確立した [16]．

結果

本研究では，統計データである国勢調査に関するデータ，科学分野のデータである抗原抗体反応に関する実験データに特異性指向マイニング技法を適用し，その有用性を確認した．ここでは，抗原抗体反応に適用したときの結果を示す [14]．

このデータは，ニワトリリゾチームを抗原とする抗体 (HyHEL-10) に関するアミノ酸配列

及び結合係数，熱力学実測データである．この実験の目的は，アミノ酸配列の変化によって結合定数のみならず熱力学特性も変化するが，これらの配列と結合係数あるいは熱力学特性との相関関係を発見することである．

表 1: 属性一覧

No.	属性名	説明
1	Amino-acid	サンプル番号
2-111	VH 1-110	抗体 VH 鎖アミノ酸配列
112-231	VL 1-120	抗体 VL 鎖アミノ酸配列
232	Temperature	測定温度
233	pH	測定 pH
234	Ka	結合定数
235	DG	自由エネルギー変化
236	DH	エンタルピー変化
237	TDS	エントロピー変化
238	DCp	比熱変化

このデータは35の実験データから成り，238個の属性を持っている．また，この属性は表1のような構成になっており，表2はそのデータの一部である．このデータの特徴は，1) 記号データと連続値が混在している．2) 属性の数が非常に多い．3) 属性の数に対し，インスタンスの数が少ない．4) 属性のほとんどが記号データである．5) 構造の変化が少なく，変化の全くない属性も存在する．6) 連続値にはミッシングデータも存在する．といったことが挙げられる．

このデータセットに $m = 0.5, \alpha = 1$ として，特異性指向データマイニング技法を適用したところ，表3のような特異データを得た．ただし，特異データが発見されなかった属性については省略した．この結果より，属性 Ka に注目すると，PF の最も高かった特異データは 42 であり，その番号は 23 である．そこで，この実験データに注目すると，ほかのデータと比べて構造の変化が全く無い．ところで，この実験の目的は，アミノ酸の構造が変化したとき熱力学特性はどのように変化するか，である．そこで，Ka の変化が最も大きなデータは，26 番目の 0.04 である．この 26 番目の実験データに注目すると，属性 DG の -32.6，DH の -53.4，DC の -0.92 は，特異データであることが分かる．また，VL アミノ酸配 32 番目の a は特異データである．ここで，Ka と DG，DH，DCp の間の相関

表 2: アミノ酸配列と実験データ (一部)

No.	VH31	VH32	...	VL31	VL32	...	Ka	DG	DH	TDS	DCp
1	*a**	d	...	n	n	...	9.6	−46.3	−97.9	−51.6	−2.25
2	s	*a*	...	n	n	...	10	−46.4	−112.9	−66.5	−2.15
⋮	⋮	⋮	...	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮
22	s	*a*	...	n	n	...	4	−44.1	−114.2	*−70.1*	−2.1
23	s	d	...	n	n	...	*42**	*−50.2*	−91.5	−41.3	−1.4
24	s	d	...	*a**	n	...	*34*	−49.5	−106.3	−56.8	−2.42
25	s	d	...	*d**	n	...	8.8	−46.1	−105.8	−59.7	−2.31
26	s	d	...	n	*a**	...	0.04	*−32.6**	*−53.4**	−20.8	*−0.92**
27	s	d	...	n	*d**	...	0.97	−40.5	−53.6	*−13.1**	−1.59
⋮	⋮	⋮	...	⋮	⋮	...	⋮	⋮	⋮	⋮	⋮

表中の斜体文字は特異データである．
また，*の付いているデータは PF が最も高かったデータである．

表 3: 発見された特異データ一覧

属性	特異データ	属性	特異データ	属性	特異データ
VH31	{a}	VL31	{a}	Ka	{42}, {34}
VH32	{a}, {e}, {n}	VL32	{a}, {d}	DG	{−50.2}, {−38, −36.4, −32.6}
VH33	{a}, {l}, {f}, {w}	VL50	{a}, {f}	DH	{−53.4}
VH50	{a}, {l}, {f}	VL53	{a}, {e}	TDS	{−70.1}, {−13.1}
VH53	{a}, {l}, {p}, {w}	VL91	{a}	DCp	{−0.92}
VH56	{a}	VL92	{a}, {d}		
VH58	{a}, {l}, {f}	VL96	{f}		
VH98	{a}				
VH99	{a}				

{} で括られている部分が，発見された特異データ (セット) である．

関係を調べると，Ka と DG の間には相関関係があることが分かる．よって，ルールの生成では DG を省略することができる．

以上より，このデータからは次の特異データを発見することができる．

アミノ酸配列の 32 番目 = a ↓ Ka の値 = 最小 DH, DC = 特異データ	Ka の値 = 最小 DH, DC = 特異データ ↓ アミノ酸配列の 32 番目 = a

これは，

- VL アミノ酸配列の 32 番目が a に変化 しているならば，
  Ka の値は最小 となり，DH, DC は特異データ となる．
- Ka の値が最小 で，DH, DC が特異データ であるならば，
  VL アミノ酸配列の 32 番目が a に変化 する．

ということを意味しており，専門的知識を用いて概念化をすることにより，より良い知識になると考えられる．

考察

データマイニングでは，新しく興味のある知識を発見することが目的であるが，これまでのような統計学を用いた方法では，このような知識を発見することは難しいものであった．しかし，特異性指向マイニングを用いて，今まで切り捨てられていた特異データに注目することで新しい知識を発見することができるようになった．

今後は，マイニングプロセスのメタレベルの制御メカニズム，複数のエージェント，マルチデータソースからのマイニング手法を開発する．また，現在準備している Web ログ・人間の認知活動に関する実験データ・医療データに対して実験を行い，さまざまな分野で有効であることを確認する [7, 8, 9, 10, 11, 12, 13, 15].

参考文献

[1] Fayyad, U.M., Piatetsky-Shapiro,G., Smyth, P., and Uthurusamy, R. (eds.) *Advances in Knowledge Discovery and Data Mining*, AAAI Press (1996).

[2] Gale, W.A. (ed.) *Artificial Intelligence and Statistics*, Addison-Wesley (1986).

[3] Matheus, C.J., Chan, P.K., & Piatetsky-Shapiro, G. "Systems for Knowledge Discovery in Databases", *IEEE Transactions on Knowledge and Data Engineering*, Vol.5, No.6 (1993) 904-913.

[4] Piatetsky-Shapiro, G. and Frawley W.J. (eds.), *Knowledge Discovery in Databases*, AAAI Press (1991).

[5] Zhong, N. "Knowledge Discovery and Data Mining", in the Encyclopedia of Microcomputers, Volume 27 (Supplement 6), Marcel Dekker (2001) 235-286.

[6] Zhong, N. and Ohsuga, S. "Automatic Knowledge Discovery in Larger Scale Knowledge-Data Bases", in C. Leondes (ed.) The Handbook of Expert Systems, Vol. 4, Academic Press (2001) 1015-1070.

[7] Zhong, N., Liu, C., and Ohsuga, S. "Dynamically Organizing KDD Process", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 15, No. 3, World Scientific (2001) 451-473.

[8] Zhong, N., Dong, J.Z., and Ohsuga, S. "Rule Discovery by Soft Induction Techniques", Neurocomputing, An International Journal, Vol. 36 (1-4) Elsevier (2001) 171-204.

[9] Zhong, N., Dong, J.Z., Liu, C., and Ohsuga, S. "A Hybrid Model for Rule Discovery in Data", Knowledge Based Systems, An International Journal, Vol 14, No. 7, Elsevier (2001) 397-412.

[10] Zhong, N., Dong, J.Z., and Ohsuga, S. "Using Rough Sets with Heuristics to Feature Selection", Journal of Intelligent Information Systems, Vol. 16, No. 3, Kluwer (2001) 199-214.

[11] Zhong, N. and Skowron, A. "A Rough Sets Based Knowledge Discovery Process", International Journal of Applied Mathematics and Computer Science, Vol. 11, No. 3, Technical University Press, Poland (2001) 101-117.

[12] Zhong, N. "Rough Sets in Knowledge Discovery and Data Mining", Journal of Japan Society for Fuzzy Theory and Systems, Vol. 13, No. 6 (2001) 581-591.

[13] Liu, C. and Zhong, N. "Rough Problem Settings for ILP Dealing with Imperfect Data", Computational Intelligence, An International Journal, Vol. 17, No. 3, Blackwell Publishers (2001) 446-459.

[14] Zhong, N., Ohshima, M., and Ohsuga, S. "Peculiarity Oriented Mining and Its Application for Knowledge Discovery in Amino-acid Data", D. Cheung, G.J. Williams, Q. Li (Eds) Advances in Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence 2035, Springer-Verlag (2001) 260-269.

[15] Wu, J. and Zhong, N. "An Investigation on Human Multi-Perception Mechanism by Cooperatively Using Psychometrics and Data Mining Techniques", Proc. 5th World Multi-Conference on Systemics, Cybernetics, and Informatics (SCI-01), in Invited Session on Multimedia Information: Managing and Processing, Vo. X (2001) 285-290.

[16] Zhong, N., Yao, Y.Y., Ohshima, M., and Ohsuga, S. "Interestingness, Peculiarity, and Multi-Database Mining", Proc. 2001 IEEE International Conference on Data Mining (IEEE ICDM'01), IEEE Computer Society Press (2001) 566-573.

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

研究代表者　松本 裕治　(奈良先端科学技術大学院大学 情報科学研究科)
研究分担者　新保 仁　(奈良先端科学技術大学院大学 情報科学研究科)
研究協力者　山田寛康　(奈良先端科学技術大学院大学 情報科学研究科)
中川哲治　(奈良先端科学技術大学院大学 情報科学研究科)
工藤拓　(奈良先端科学技術大学院大学 情報科学研究科)
山本薫　(奈良先端科学技術大学院大学 情報科学研究科)

背景と目的

医学生物学分野のような専門性の高く，かつ，大容量の文献データベースに対し，利用者の要求に応じた検索を行うこと，あるいは，利用者にとって重要な情報を抽出することは重要な技術である．このような論文データには，専門用語が頻出するが，多くの言語処理システムにとって，それらは未知語であることが多い．以前の我々の調査では，Medline アブストラクト[1]の論文要旨には，通常の辞書には含まれない語が約 15%含まれていた [6]．また，専門用語は造語性が高く，新しい語を辞書に登録するという作業を続けても，未知語の問題を完全に回避することはできない．日々更新される論文データから有用な文献の検索，あるいは，有用な情報の抽出を行うための言語処理を考えると，その基本となるのが，専門用語の同定およびその意味クラス分類である．

本研究では，ある特定の意味クラスに属する用語の発見を目的とし，論文要旨に出現する名詞句がそのクラスの用語であるかどうかを同定するタスクとして問題設定を簡略化し，学習に基づく手法の性能について実験結果を報告する．特に，学習事例が少ないという現実的な想定の基に，用語が出現するテキスト中の文脈が意味クラス推定にどの程度有用であるかを明らかにすることを目指した．また，用語の内部情報として，用語の主辞となる語の部分文字列情報の有効性についても検討するため，この情報を用いる場合と用いない場合の比較実験も行ったので，合わせて報告する．

検討内容

本年度の研究項目

医学生物関係の論文など専門性の高い分野の文献に対して言語処理を行う際の問題は、多くの単語が既存の辞書には登録されていないため、言語解析システムにとって未知語になってしまうことである。そのため、未知語の存在を想定した言語処理が必要である。また、文書中から専門用語を同定するために名詞句等のまとめあげを行うことが必要である。

このような状況を考慮し、本年度は、以下のような基礎的な言語処理システムの構築と、専門用語の意味クラス推定に関する実験を行った。

1. 未知語を含む英文中の単語の品詞推定

[1] http://www.ncbi.nlm.nih.gov/PubMed

表 1: 英語の品詞付与の精度 (既知語/未知語)

Training Tokens	SVMs				TnT
	Preceding POS	d	Preceding & Succeeding POS	d	
1,000	83.4%(96.3/68.3)	1	83.9%(96.4/69.3)	1	83.8%(96.0/69.4)
10,000	92.1%(95.5/81.2)	2	92.5%(95.7/82.2)	2	92.3%(95.7/81.5)
100,000	95.6%(96.5/85.7)	2	95.9%(96.7/86.7)	2	95.4%(96.4/83.3)
1,000,000	97.0%(97.3/86.3)	2	97.1%(97.3/86.9)	2	96.6%(96.9/84.2)

2. 文書中の名詞句等の基本句の自動抽出

3. 専門用語の意味クラス推定

なお、これらの研究は、蓄積された正しいデータからの学習に基づくシステムの構築を目指し、学習手法として Support Vector Machine を利用した。本報告では、未知語の品詞推定と基本句抽出については、実験結果の報告に留め、主として、3 点目の専門用語の意味クラス推定について報告する。

未知語を含む英文文書内の単語の品詞推定

専門分野の文書には一般の辞書には含まれない語が多く出現し、品詞等の文法情報の特定に支障をきたす。専門性の高い分野には次々に新しい用語が出現するため、それらをすべて辞書に登録することは現実的に不可能である。そこで、前後の文脈、あるいは、単語の綴り (特に接頭、接尾表現) を手がかりとして、未知の単語の品詞を決定し、それによって、専門用語と考えられる名詞句の同定を柔軟に行うことを試みた [12][14]。

詳細は、省略するが、前後の品詞情報や対象語の綴りの一部を属性情報とし、Support Vector Machine[20] を利用した英語の品詞判定手法を考案し、実験を行った。表 1 に学習データの量を変更した場合の提案システムの品詞付与精度を示す。参考として、最後の欄に、統計的品詞付与システムとしてよく知られている TnT[2] の精度を示す。

なお、本システムで利用した SVM は、このように高い学習能力を有するものの、学習および実行に時間がかかり、実用面では問題があった。そのため、第一の処理として、従来型の N-gram に基づく統計的品詞付与システムによる学習を行い、そのシステムが誤りを生じる箇所を SVM によって学習するという二段構えの方法を提案し、修正学習と名付けた [13]。表 2 に、品詞 trigram モデルにより学習を行ったシステムの精度 (T3 Original で示めされた行)、修正学習の精度、TnT の精度、および、前記のすべて SVM で学習を行った場合の精度を示す。修正学習においては、訓練データの全ての単語や品詞情報を使った場合と、未知語の性質を学習させるため出現頻度 1 回の単語を取り去った場合の精度を比較した。後者の方が、未知語に対する精度が改善されていることがわかる。

この実験から、SVM による修正学習によって現状の統計的品詞付与システムを上回る精度を達成できることがわかった。ただし、すべての部分を SVM によって学習した場合

表 2: 修正学習による実験結果と他手法との比較

	精度 (既知語 / 未知語)	誤り数
T3 Original	96.7% (96.9% / 82.7%)	9720
修正学習	96.9% (97.2% / 83.6%)	8734
修正学習 (cutoff-1)	97.0% (97.3% / 85.1%)	8588
TnT	96.6% (96.9% / 84.2%)	9626
純粋 SVM	97.1% (97.3% / 86.8%)	8245

表 3: Computational Cost on WSJ Corpus

	訓練事例数	学習時間 (時間)	テスト時間 (秒)	精度
T3 Original	—	—	89	96.59%
修正学習 2 次多項式 kernel)	1027840	16	2089	96.98%
修正学習 1 次多項式 kernel)	1027840	2	129	96.94%
純粋 SVM	999984×50	625	55239	97.11%

に比べると、やや精度が下回ることがわかった。訓練時間や解析時間についても調べたところ、表 3 のような結果を得た。SVM の学習には多項式 kernel を用いたが、線形 kernel を用いたところ、学習時間、解析時間とも問題ない範囲で実行することができた。比較として、すべての学習を SVM を用いた場合を最後の行に示した。解析精度はわずかに上回るものの、学習時間で約 300 倍、解析時間で 400 倍以上の差があることがわかった。1 次の多項式 kernel を用いた場合は、trigram モデルと比較しても、1.5 倍程度の時間しか要しないことがわかった。

テキストからの専門用語の抽出と分類

専門用語抽出・分類の自動化

テキストからの専門用語の抽出とその意味クラス同定については，人手による規則に基づく方法と計算機による自動化に基づく方法の 2 つのアプローチに大別することができる．用語の抽出については，ある程度の共通の規則の設定が可能であるが，用語の意味クラスの推定 (分類) については，現段階では，人手に基づく手法が精度の点で高い数値を挙げている．しかし，異なる意味クラス毎に規則を用意する必要があること，異なる分野によって専門用語の形式が依存することを考えると，規則の記述およびその更新に要するコストは見逃すことができない．一方，計算機による専門用語の意味クラス推定については，専門用語のタグ付けを行ったテキストを用意し，そこから規則を自動的に学習するという手法がとられることが多い．この方法では，規則の記述そのものについてのコストは

かからないが，タグ付きデータの蓄積，および，学習に用いる素性としてどのような情報を利用すべきかを考察することが利用者にとっての負荷となる．

前者の人手による抽出については，例えば，福田ら [4] は，タンパク質名とそれに関係する物質名の抽出を行う規則を人手によって記述している．この方法では，タンパク質名に特有な単語を core-term として正規表現等によって定義し，その前後の単語を規則によって連接することにより，複合語としての用語の抽出を行っている．その結果，95%程度の高い正解率を達成している．

一方，計算機による学習による方法については，合原ら [6] は，医学生物学分野の文献から 13 種類に分類した専門用語の自動抽出と分類実験を行っている．彼らは，単語の文字種，品詞，用語に関する係り受け情報を素性として，決定木学習による実験を行っている．また，co-training[1] を利用し，少量の訓練データに合わせて大量の未知データの利用の有効性を確認しているが，co-training による効果は限定されたものとなっている．Collier[3] らは，隠れマルコフモデルを用いたゲノム関連用語の抽出とクラス分けの実験を報告している．専門用語のクラス分けには，他の単語との共起や文字列などの細かい素性が必要となる．そこで我々は，以前，Support Vector Machine[20] を用いて，単語や部分文字列などの詳細な情報を用いた高次元の素性空間の下で，専門用語の抽出とクラス推定を試みた [19]．部分文字列の利用の有効性を確認することはできたものの，用語のクラス分類については，高い性能を達成することができなかった．その大きな原因は，学習データの不足および用語自体が持つ曖昧性によるものと考えている．特に，用語の曖昧性については，単一の語がタンパク質名を指したり，DNA の配列を指したりすることがあり，同定が容易でない場合がある．我々のデータでも，人手によるタグ付けで 20%程度の不一致を観測した．同程度の揺れについては，太田ら [16] によっても同様の結果が報告されている．

専門用語抽出について

専門用語の抽出について問題となるのは，専門用語がどのような文法的構造をもっているかという点と，それを一般の用語とどのように区別するかということである．これについて，Justeson[7] による興味深い報告がある．3 種類の分野における彼らの調査によると，専門用語辞書に現れる用語のうち 92.5%〜99%が文法的には名詞句であり，2 単語以上から構成される用語のうち 97%が単純な名詞句 (形容詞と名詞のみからなる) である．専門用語とそれ以外の名詞句は語彙的名詞句 (lexical NP)，非語彙的名詞句 (nonlexical NP) と言う呼び方で区別され，前者はそれ自体が独特の意味を持ち，辞書に登録すべき語と見なされている．その語彙的な性質から，その語がテキスト中で再び参照される場合には，非語彙的名詞句とは異なり，主辞あるいはより一般的な名詞によって省略された形ではなく，その語形全体を引用して参照されやすいと議論している．この考察に基づき，テキスト中に 2 度以上現れる複合名詞句を抽出することにより高い精度の用語抽出が実現できることが示されている．ただし，この考察では，対象とするテキストは論文等の比較的長いテキストであり，アブストラクトなどの短いテキストでは，複数出現の条件は厳し過ぎると考えられる．

専門用語抽出と意味クラス推定実験の準備について

本稿では，Justesonらの考察のうち，専門用語のほとんどが文法的には名詞句であることに従い，名詞句を対象として，その意味クラス分類に焦点をあてる．その語が，専門用語としてその意味クラスに属するのか一般語として属するのかの区別は，ここでは対象外とする．また，対象として，特定の意味クラス (本稿では病名) に限って実験を行った．前提とした考え方は，以下の通りである．

- 用語の意味クラスの同定と，用語が辞書にそのまま登録すべき専門用語であるかどうかというタスクは，独立の問題であること．後者については，Su[18], Maynard[11]などの研究がある．
- テキストからの名詞句については，汎用のシステムでかなり高い精度で自動抽出が可能である．
- 用語の意味クラス推定を行うために，人手によってタグ付けされた大量のテキストを仮定することは現実的でない．

以上の考察に基づき，本稿では，Medlineアブストラクトを題材とし，テキスト中に現れる名詞句を対象にして，その意味クラスを推定するタスクに関する実験を行った．論文中には多くの語が言語解析システムにとっては未知語として出現するが，これには我々のグループで行った英語の未知語推定のシステム [12]，および，名詞句チャンキングのシステム [8, 9] を利用した．前者の未知語推定は，品詞タグ付け処理の一環として行われ，辞書に現れない未知語については，その語尾，語頭および文字種の情報と前後の単語や品詞情報用いて，品詞推定を行っている．未知語に対する品詞の推定精度は約87%であり，全体としての品詞タグ付け精度は約96%である．後者の名詞句の推定は，名詞句だけでなく，動詞句，形容詞句，副詞句，前置詞句など10種類の基本句のタグ付けを行う一般的なシステムとして構築されたものを用いている．名詞句の同定の精度は約95%である．

以前の我々の研究 [6][19] では，人手によって専門用語の箇所と意味クラスのタグ付けが施されたコーパスを実験に用いた．人手によるコストが大きいため，約1500事例 (タグ付けの総数) による小規模な実験しか行えなかった．今回は，ある程度教師なしの手法を実践することを目指し，コーパスの精度は落ちるものの，次のように人手による作業をなるべく少なくしつつ，より大規模なデータに対して，特定の意味クラスの語の推定実験を行った．

1. 医学生物関係の日英対訳辞書 [15] から，病名を表す用語を抽出: 辞書には病名かどうかの記述がなかったため，明らかに病名と思える日本語訳 (〜病，〜症，など) をもつ英語の用語を抽出 (約3000語)．これをターゲット用語の一覧とする．
2. これらの病名を用いてPubMedを検索し，各1個のアブストラクトを得る．
3. 得られたアブストラクトの品詞タグ付け (未知語処理を含む) と名詞句チャンキングを上記ツールにより解析し，すべての名詞句を抽出する．

アブストラクト数	ターゲット (病名用語数)	非ターゲット (反事例数)
508	845	7189
1272	3350	17033
1907	6150	25689

表 4: 病名およびそれ以外の用語の統計

4. 上記の病名以外の名詞句をランダムに取り出し，病名以外のものを人手により判定し抽出．これをターゲット用語の反例とする．

5. テキスト中の名詞句に対し，ターゲット用語/非ターゲット用語/それ以外，の 3 種類のタグを付与する．

これらの処理のうち，1. および 4. の処理には人手が介入が必要である．1. の処理では，英語名称に対する限定は行っていないが，元となる日本語名称に対する制限を行っているため，病名の選択に偏りが生じている可能性がある．また，4. の処理では，明らかに病名でないと思われる用語を頻度順に取り出した．さらに，5. では，これらの用語を実際のテキストと照合する際に，テキスト中の名詞句との完全な一致ではなく，名詞句の主辞 (中心語，基本的には末尾の名詞) の活用処理を行って原型に直し，さらに名詞句全体が一致しなくても，名詞句の末尾部分が用語と完全に一致した場合にも，その名詞句を一致した用語と同じクラスとみなすことにした．

論文アブストラクトの数に対応したターゲット用語 (病名)，反例の数を表 4 にまとめる．これらは，学習時の正例および負例に対応する．表からわかるように，単純な抽出にもかかわらず，病名である正例は負例データの数に比べて 1 桁近く少ない．

実験と考察

用語の意味クラスを同定するために必要な情報は，用語の内部情報と文脈情報に大別することができる．用語の内部情報は，用語を構成する単語および単語の綴りに関する情報である．具体的に，“〜 disease” のように末尾の単語を見るだけで明らかに病名とわかる用語もあれば，“hepatitis(肝炎)” のように単語の末尾の綴りからそれが病名であると予測できるものもある．一方，文脈情報は，用語の使われ方に関する情報であり，用語が文内の周辺単語とどのように関係するかによって，その語の意味クラスを推定しようとするものである．本稿では，用語のバリエーションがある程度限定されてしまっていると考えられるので，用語の内部情報を過度に利用するのは避けることにし，文脈情報によってどの程度用語の意味クラス推定が可能かを確認することを主たる目的とした．

実験の手順と結果

以下の実験では，SVM(Support Vector Machines)[2]を用い，用語の候補とその前後の文脈情報を利用した学習を行った．用語の候補となるのは，論文アブストラクに含まれる名

[2]実験には，TinySVM[10] パッケージを使用した．

詞句で，かつ，先に抽出しておいたターゲット語および非ターゲット語である．

ターゲット語である 3000 用語から，ランダムに約 500 個，1200 個，および，1900 語を選択し，規模の異なる 3 通りの実験を行った．それぞれの実験では，各ターゲット語を用いて PubMed を検索し，得られた論文アブストラクトから，ターゲット語および非ターゲット語を抽出した．したがって，ターゲット語は，必ず 1 度は出現したことになる．各用語 (ターゲットおよび非ターゲット) を含む文の品詞タグ付けとチャンキングを行った．用語は，今回の実験では，必ず名詞句となっている (用語を含む名詞句を改めて「用語」とした)．学習およびテストには，対象となる用語となっている名詞句の前後 2 つのチャンク (句) を文脈情報として利用し，各チャンクの意味的中心語である主辞となっている単語とその品詞タグを用いた．文脈情報としては，対象用語に文法的に直接係っているチャンク，あるいは，対象用語に直接係られているチャンクを選ぶべきであるが，現在入手可能な英語の係り受け解析の精度の関係で，今回の実験では，係り受け解析の方法は利用しなかった．

個々のサイズのデータを 5 等分して，交差検定[3]により，未知の用語に対する意味クラス推定実験を行い，学習の精度と再現率を測った．試験は，学習結果の SVM を用いて，与えられた用語が病名か否かを識別する分類問題となる．なお，精度は，学習システムが病名と判断した用語の正解率であり，再現率は，本来病名である用語のうちシステムが病名と判断したものの割合を指す．

学習およびテストに用いた情報 (素性) をまとめると，次の通りである．

- 用語の前後 2 つのチャンクのラベル
- 用語の前後 2 つのチャンクの主辞の単語
- 用語の前後 2 つのチャンクの主辞の品詞名
- 用語の主辞の単語の末尾 3 文字および 4 文字 (内部情報を用いた実験でのみ使用)

表 5 に文脈情報のみを用いた場合の学習システムの精度と再現率を示す．3 つの列は，SVM で用いた多項式カーネルの次数を表す．これは直観的には，学習システムが素性のいくつまでの組合せを考慮したか，に対応する．次数 1 の多項式，すなわち，線形関数は，予備実験の段階で他のカーネルに比べて精度，再現率とも劣ったので，詳細な実験は行わなかった．表では，5 回の交差検定のそれぞれの結果とその平均を示した．

実験結果に対する考察

これらの結果から，学習データの増加に伴って全体的な性能が向上していることがわかる．また，学習事例が少ない場合には，3 次の多項式カーネルを用いた場合が，精度と再現率のバランスがよいと言えるが，学習事例が多くなると，カーネル関数間の差がほとんどなくなるようである．なお，事例の増加は，精度に比べて再現率により効果を与えている．

[3] 5 等分したデータの 4 つを学習データとし，残りの 1 つを試験データとして精度と再現率を測る．これを異なる組合せにより 5 回の実験を行った．

アブストラクト数	多項式カーネルの次数		
	2	3	4
508	47.4/30.7	52.4/30.2	58.6/28.5
	46.7/32.5	53.3/32.5	56.1/30.5
	49.5/29.8	57.8/27.0	58.6/23.0
	47.3/32.7	52.2/30.2	58.8/29.6
	53.2/37.6	60.6/37.1	61.1/32.6
平均	48.8/32.7	55.3/31.4	58.6/28.8
1272	62.0/48.9	67.2/47.7	67.4/44.7
	58.9/52.5	63.8/52.7	65.5/51.4
	60.3/48.5	65.3/47.4	68.1/46.4
	62.0/53.1	66.3/49.9	68.4/47.6
	64.2/52.1	68.9/49.2	69.8/46.5
平均	61.5/52.0	66.3/49.4	67.8/47.3
1907	68.8/59.4	67.5/58.7	69.2/56.7
	72.6/58.3	72.0/59.1	73.9/57.3
	66.5/57.6	67.8/58.9	68.9/56.9
	71.0/59.2	69.2/60.0	70.2/57.6
	67.3/56.4	66.4/56.7	68.1/55.1
平均	69.2/58.2	68.6/58.7	70.1/56.7

表 5: 文脈情報のみによる実験の結果

多項式カーネル次数の増加は，精度には好影響を与えるが，再現率にはむしろ悪影響であることがわかる．これは，素性の組合せを細かく考えることにより，より制限的な学習が起こっていると考えることができる．用語の内部情報を利用せず，文脈情報だけを用いた病名推定は，精度が 6 割から 7 割，再現率が約 5 割から 6 割と，あまり高性能の結果が得られていない．ただし，病名およびそれ以外の用語の比率は，11%〜20%であり，ランダムな推定を行うことと比較すると遥かによい結果であると言える．

今回の実験では，用語の種類が限られているため，用語の内部情報を利用するのは，本当の意味ではフェアではない．表 6 は，文脈素性に加えて，対象用語の主辞の末尾の文字列 (長さ 3 文字および 4 文字) を素性として利用した実験結果であり，参考のために示した．これによると，用語の主辞単語の末尾 3，4 文字列を用いることで，精度，再現率とも大幅に向上していることがわかる．

全体的な考察

本稿では，大量のテキストにタグ付け作業を人手によって行うことなく，専門用語の意味クラス推定を自動化することを目指して，SVM を用いた学習に基づく手法の実験結果を示した．今回の実験における問題点を改良するためのいくつかの可能性をここで考察し

アブストラクト数	多項式カーネルの次数		
	2	3	4
508	86.1/76.0	87.6/63.1	90.9/55.9
	86.7/68.9	89.9/58.9	91.8/51.7
	82.8/62.4	83.7/55.1	82.9/48.9
	90.5/66.0	90.0/56.6	88.2/47.1
	89.2/69.7	92.6/62.9	94.4/57.3
平均	87.1/68.6	88.8/59.3	89.6/52.2
1272	92.1/81.4	92.3/79.0	93.4/73.0
	91.0/82.0	92.4/79.8	93.2/75.4
	88.8/82.2	91.5/78.3	92.5/73.5
	90.4/80.0	90.1/76.3	90.4/71.1
	91.8/79.1	94.3/75.8	95.7/71.6
平均	90.8/80.9	92.1/77.8	93.0/72.9
1907	91.8/85.4	91.7/82.7	92.3/80.0
	92.5/82.7	93.0/80.6	92.8/77.1
	92.1/83.4	92.8/82.1	93.3/79.1
	90.9/83.1	91.7/82.1	92.6/78.7
	92.3/84.5	91.7/82.8	91.8/79.0
平均	91.9/83.8	92.2/82.1	92.6/78.8

表 6: 文脈情報と文字列情報による実験の結果

たい.

一つは，より精度の高い言語処理の適用である．今回用いた未知語処理を伴う品詞推定，および，句へのチャンキングプログラムは，現在発表されているシステムの中では最も高い精度を示しており，現時点では充分優れたものである．しかし，それぞれのシステムは，現在入手可能なタグ付きコーパスである Penn Treebank[4]から学習を行ったものであり，今回利用した医学生物学分野の論文とは，内容が大きく異なるものである．今後，類似分野のタグ付きコーパスを蓄積することによって，より精度の高い解析を行える可能性がある．また，今回は利用できなかったが，英語の単語あるいは句単位の係り受けに対しても統計的手法がいくつか提案されており，係り受けレベルで 90%程度の解析精度が可能であることが示されている．今後，このような高精度な言語処理システムを用いた実験を行うことが必要であると考える。

Justeson ら [7] の考察にもある通り，重要な専門用語は，文献中で複数回出現することが多いと考えられる．また，同じ用語は，同一テキスト内では同じ意味で用いられるのが普通であり，この性質は，単語の語義の曖昧性の解消の研究でも利用される性質である

[4] http://www.ldc.upenn.edu/

[5]．今回の実験では，用語の個々の出現について，その意味クラスを同定することを試みた．しかし，同一文献に同じ用語が複数回出現する場合には，それらの文脈情報をすべて利用して意味クラスを推定することも考えられる．

学習に Support Vector Machine を分類器として用いたことから，病名の推定を行うために，病名の一覧だけでなく，負例として病名でない専門用語の集合が必要となった．今回は，負例の選択にある程度人手のコストをかけた上，負例の抽出は極めて不完全なものにならざるを得なかった．あるクラスの用語を抽出するための学習を行う際に，正例の蓄積は意味がある作業だが，負例の抽出は必ずしも意味があるとは言えない．SVM の一つに，One-class SVM[17] といって，正例だけを仮定する学習法がある．今後，このような学習を用語の意味クラス同定にも適用することを考えたい．

参考文献

[1] Blum, A. and Mitchel, T., “Combining Labbeled and Unlabeled Data with Co-training,” COLT98, pp.92–100, 1998.

[2] Brants, Thorsten, “TnT — A Statistical Part-of-Speech Tagger,” Proceedings of the 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association of Computational Linguistics, pp.224-231, 2000.

[3] Collier, N., Nobata, C., Jun-ichi Tsujii, “Extracting the Names of Genes and Gene Products with a Hidden Markov Model,” COLING2000, pp.201-207, 2000.

[4] 福田賢一郎, 他 “医学生物学文献からの専門用語の抽出に向けて,” 情報処理学会論文誌, Vol.39, No.8, pp.2421-2430, 1998.

[5] Gale, W., Churck, K., Yarowsky, D., “One Sence per Discourse,” Proc. Speech and Natural Language Workshop, pp.233-237, 1992.

[6] 合原博, 宮田高志, 松本裕治, “医学生物学分野からの専門用語抽出・分類,” 情報処理学会 自然言語処理研究会報告, 2000-NL-135, pp.41-48, 2000.

[7] Justeson, J.S. and Katz, S.M., “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text,” *Natural Language Engineering*, Vol.1, Part 1, pp.9-27, 1995.

[8] Kudo. T. and Matsumoto, Y., “Use of Support Vector Machines for Chunk Identification,” *CoNLL2000*, pp.142-144, 2000.

[9] Kudo. T. and Matsumoto, Y., “Chunking with Support Vector Machines,” *NAACL2001*, pp.192-199, 2001.

[10] Kudo, T., “TinySVM: Support Vector Machines,” Nara Institute of Science and Technology.
`cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/`

[11] Maynard, D. and Ananiadou, S., “TRUCKS: A Model for Automatic Multi-Word Term Recognition,” 自然言語処理, Vol.8, No.1, pp.101-125, 2001.

[12] 中川哲治, 工藤拓, 松本裕治, “Suppoert Vector Machineを用いた未知語の品詞推定,” 情報処理学会 自然言語処理研究会報告, 2001-NL-141, pp.77-82, 2001.

[13] 中川哲治、工藤拓、松本裕治, 「修正学習法による形態素解析」, 情報処理学会研究報告 2001-NL-146, pp.1–8, 2001 年 11 月.

[14] Tetsuji Nakagawa, Taku Kudoh, Yuji Matsumoto, “Unknown Word Guessing and Part-of-Speech Tagging Using Support Vector Machines,” Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS2001), Tokyo, pp.325–331, November 2001.

[15] 日外アソシエーツ, “バイオ・メディカル用語対訳辞典 (英和・和英),” 日外アソシエーツ株式会社, 1999.

[16] 太田朋子, 他, “医学・生物学論文からのタグ付きコーパスの作成,” 情報処理学会 自然言語処理研究会報告, 1999-NL-133, pp.93-98, 1999.

[17] Scholkopf, B., et al., “Estimation the Support of High-dimensional Distribution,” *Technical Report MSR-TR-99-87*, Microsoft Research, 1999.

[18] Su, Keh-Yih, Wu, Ming-Wen and Chang, Jing-Shin, “A Corpus-based Approach to Automatic Compound Extraction,” *ACL'94*, pp.242-247, 1994.

[19] 山田寛康, 工藤拓, 松本裕治, “単語の部分文字列を考慮した専門用語抽出と分類,” 情報処理学会 自然言語処理研究会報告, 2000-NL-140, pp.77-84, 2000.

[20] Vapnik, V.: Statistical Learning Theory, *John Wiley & Sons*, 1998.

A03 班：アクティブユーザリアクション

アクティブマイニングと EBM

研究分担者　津本周作　(島根医科大学医学部医療情報学講座)
研究協力者　平野章二　(島根医科大学医学部医療情報学講座)
高林 克日己　(千葉大学医学部附属病院医療情報部)
柳樂 真佐実　(島根医科大学医学部医療情報学講座)

背景

計算機の能力の向上とデータベースソフトウェアの性能の向上により，遺伝子データベース，癌プロトコルデータベースを始めとした医学データベースのみならず，検査データベース，検診データベースを含めた診療データベースも膨大な量のデータが蓄積されるようになってきた．このような膨大な量のデータの解析はすでに人の処理能力をはるかに越えるものとなっており，計算機による有効な使用方法の確立が医学・医療の分野においても急務とされている．医学においてはデータ解析が統計学的手法を用いることが伝統的であり，その総決算ともいえるのが，最近強く叫ばれるようになったのが EBM(Evidence-based medicine) である．EBM を支える基本的な概念は “Statistical Evidence” と呼ばれるものであるが，この Statistical Evidence を確立するためには，非常に精密な実験計画と実験的なデータ収集が必要である．現在蓄積されつつある電子化データが必ずしも EBM で求められる最良の Evidence を確立するための条件を満たしていないが，EBM の実践の過程を反省すれば，電子化データから Statistical Evidence に至るまでに，アクティブマイニングが指向しているアクティブ情報収集，ユーザ指向アクティブマイニング，アクティブユーザリアクションが大きな役割で果たすであろうことが展望できる．本稿では，この点に着目し，アクティブマイニングと EBM との関係について論じる．

EBM とは

EBM の背景

従来の医療は患者が異なり，診療する医師が異なれば，検査や治療の進め方が異なるのは当然であり，治療に関する意思決定は個々の医師の自由裁量によるものであった．このため,「名医」なる者が存在し，外科手術の成績が地域で数倍異なることはよく見られた．診療特に早期診断および治療は各医師の経験が大きな因子となり，患者は受診した医師によって，疾患の予後が変わることがあった．

一方，がん検診などの普及によって，信頼できるデータに基づく，スクリニーングの必要性が叫ばれるようになってきている．例えば，便潜血陽性患者の 50 この検査の陽性から大腸内視鏡検査という一つの診断フローによって，大腸ガンの早期診断が極めて有効に行われてるようになってきた．また，生活週間病を含めた有病率の高い疾患に関しては，診療に関するデータが蓄積されるにつれ，その診断および治療方法が確立，マニュアルあるいはガイドライン化化されはじめている．例えば，気管支喘息の軽症間欠型は，週 1 〜 2 回の発作で症状は間欠的で短いタイプとして定義され，治療法としては吸入/経口薬を頓用することが喘息治療のガイドラインとして与えられている．ガイドラインは統一した疾患概念に基づいて，診断・治療の標準化を行うものであり，ガイドライン策定

に関しては，多くの臨床研究を整理して，最良の診療行為を決定することが必要であり，EBM に基づいた最良のガイドライン作成がアメリカ合衆国を中心に行われはじめている (http://www.guidelines.gov)．

以上のごとく，医療は従来，偶然性の強い個人的な経験や観察に基づくことが多かったが，体系的に観察・収集されたデータに基づく医療への転換が欧米および日本を中心にはじまっており，基礎医学的な知識の臨床応用という立場から，患者から得られた観察データを重視する立場が臨床医学の常識となりつつある．これに伴い，データに対する生物統計学的手法の実践のみならず，データの収集から客観性を持たせようという努力がはじまっている．従来，このような統計学的なアプローチは疫学として研究されてきたが，このアプローチを臨床医学の上に展開したのが，EBM(Evidence-based Medicine) である．

EBM の定義

EBM はその分野での中心的メンバーである David L. Sacket らによれば，“ the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients” (入手可能で最良の科学的根拠を把握した上で，個々の患者に特有の臨床状況と価値観に配慮した医療を行うための行動指針) と定義されている [2]．これは前小節で論じたごとく，データを強く意識した医療の実践を強く意識したものである．しかし，実際に EBM の実践法に関する書物を見れば，医師が実際の治療の際に行っている手法とあまり変わりがないと感じる方も多いと思われる．これは，EBM がそれぞれの持つ独自の手法を重視しているわけではなく，様々な手法を「統合」し，その統合した結果から，適用分野に関して最良の結論を導出することに重点を置いており，プロセスの統合が重要であることを指摘している点にある．この特徴は，アクティブマイニングが情報収集，解析，解釈の統合的なプロセスとしてとらえられることと共通している．

疫学から見たデータ解析

疫学・公衆衛生学ではこれまでデータ収集方法に関する知見を重視して，データ収集の方法からデータ解析を分類してきた．この視点から考察すれば，病院情報システムから抽出された医療データに限らず，データマイニングで使用されている大半のデータは一般に，ある仮説に基づいて収集されたデータではない．データマイニングはとりあえず収集されたデータから有効な情報を抽出しようというのがその目的であると論じられてきた．したがって，自ずとそのデータの説明力には限界がある．このようなデータの性質の違いは，疫学において論じられており [1]，ある仮説に基づくデータ収集を元にしたデータ解析は prospective study(前向き研究)，とりあえず集められたデータを解析することを retrospective study(後向き研究) と呼ばれてきた．Prospective study は疫学では cohort study とも呼ばれ，観察開始時点 (これはデータ収集開始時点に相当する) に対象の集団をいくつかの群に分類する (要因 (+),(-))．その後，疾病の発生や疾病による死亡等を観察していくという方法であり，要因の有無を固定した上で，それによる結果の有無を観測する方法である．この情報収集の流れは，原因 →結果の因果律にそっており，前向き研究 (prospective study) と呼ばれている．

他方，Retrospective study は case-control study とも呼ばれ，疾病の発生した人を症例

(case: positive examples に相当する) に，症例に該当しない人を対照 (control: negative examples に相当する) として，両群の要因の有無について比較する方法である．この情報収集の流れは，結果 → 原因で因果律の逆向きとなり，後ろ向き研究 (retrospective study) と呼ばれている．これら二つの方法の長所と欠点に関して，表 1 に示した．現在データマイニングで用いられているデータベースはほとんど retrospective study に相当し，一般に prospective なデータ収集によらないため，収集時におけるノイズと bias の問題が発生する．仮説検定で最終的に信頼の高い結果を得るには，やはり prospective study の方が適している．両者の特徴から考えれば，収集されたデータから retrospective analysis で仮説を生成 (**ユーザ指向アクティブマイニング**) し，その仮説に基づき，データをアクティブに収集した後 (**アクティブ情報収集**)，prospective analysis，特に統計的解析を試行するのが，データマイニングの手法を科学的な分野に適用する最良の形式であろう．有用な仮説を生成するためには，領域特有な知識による検証 (**アクティブユーザリアクション**) が不可欠である．

表 1: Prospective Study(前向き研究) と Retrospective(後向き研究) の比較

	Prospective Study	Retrospective Study
別称	Cohort Study	Case-Control Study
データ収集の流れ	要因→結果	結果→要因
推定能力	様々な指標を正確に推定できる.	正確さはデータからは推測することが困難
要因の解析能力	まれな要因でも分析可能	まれな要因についての分析は困難
解析の多重性	一つの要因に関して 複数の Target を分析可能	一つの Target に関して 複数の要因を分析可能
ノイズの影響	少ない	大きい
相対リスク	発生率の高い Target の 相対 risk の計算も可能	相対 risk の計算は不可能
調査コスト	調査対象の選別等，コストが高い	コストは低い
まれな疾患 (Target)	まれな Target の調査は困難	まれな Target の調査も可能
調査期間	Target の発生の有無の確認が必要	調査期間は短くてよい.
追跡調査 (Follow-Up)	追跡不能例の発生 (positive/negative が明確でないことあり．)	Target の positive/negative は データ収集時に確定

EBM 実践の手順

EBM でもっとも重要なことは，最良の科学的根拠を把握する手段である．もっとも望むべきは，個々の臨床医が最良の科学的根拠となるデータ解析を行えればよいのであるが，実際には EBM で推奨されるデータ解析を行うことは非常に難しい．したがって，一般には，学術雑誌に掲載されている一般化可能な手法で収集されたデータを統計学的に解析して得られた知識 (Statistical Evidence) に関して，妥当性を検討しながら，Evidence を抽出されることになる．文献検索を中心とした EBM の手順としては，図 1 の形が考えられる．例えば，痛風発作の経験のない高尿酸血の患者の症例の治療を行う場合について，EBM の実践を考えてみる [2]．痛風発作の発症機序は，高尿酸血症から関節内に尿酸が蓄積・析出，これが関節炎つまり痛風発作が発症させるとされている．では，無症候性高尿酸血症に予防治療すると，痛風発作の発生が減るだろうか? (**Step1: 疑問の定式化**) ここで，無症候性高尿酸血症の治療に関する文献を収集する (**Step2: Evidence の収集 →文献収集**)．文献が 1 件 (Cohort Study) あったとして，その文献の Evidence の質を評価する

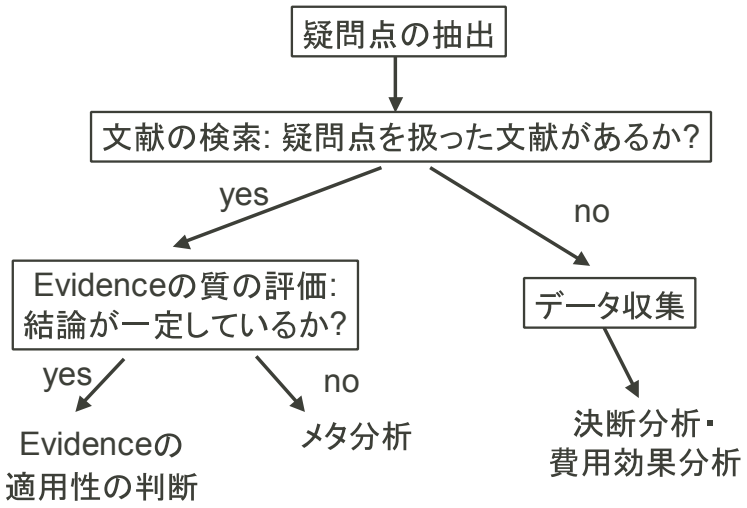


図 1: EBM の手順

(**Step3: Evidence の質の評価**)．文献における研究の結果は表 3 の通りであったと仮定する．この文献はコホート研究であるので，次の節で述べるように，Evidence としては

表 2: 収集された Evidence

	年間発症率	累積 (5 年)
$UA > 9$	0.049	0.22
$7.0 < UA < 8.9$	0.005	0.03
$UA < 7.0$	0.001	0.005

b に当たる．さて，最後に当該患者にこの Evidence の適用を考える (**Step4: 適用性判断**)．当該患者は $UA > 9$ であるので，予防治療の必要性があり，痛風としての治療を開始するべきであると判断できる．

もし，Evidence として採用できる文献がなければ，当該診療施設でのデータを収集し，そこから痛風の治療について，データから考察することが必要である (Local Evidence による EBM の実践)．この部分は，ユーザ指向アクティブマイニングに相当する．

以上のように，EBM の実践では，Evidence の収集および Evidence の質の評価 (解析・解釈) が重要視されている．

Evidence の評価

EBM では，その Evidence の質を評価する際に，各論文で使われているデータ収集の方法と統計学的検定に関して注目する．特に，データ収集の方法が，論文の結果が一般化できるかどうかの極めて重要な判断材料となる．このデータ収集の方法を EBM では，“研究デザイン” と呼び，一般化できるかどうかのランク付けを “証明力” と呼ぶ．研究デ

ザインには，ランダム化比較試験，コホート研究 (前向き研究)，症例対照研究，横断研究 (この二つは後ろ向き研究)，症例報告があり，この順で証明力が落ちるとされている．

ランダム化比較試験とは，前向き研究において，適格症例を無作為に治療群と対照群とに割り付ける方法であり，試験群比較性が優れているのが特徴である．さらに，ある危険率 α と検出力 β が与えられると，検定に必要な症例数が理論的に得られている．例えば，2 群 Event で，発生確率がそれぞれ 0.4, 0.45 の場合，$\alpha = 0.01$，$\beta = 0.95$ の場合，3486 例必要であり，発生確率がそれぞれ 0.3，0.7 の場合，$\alpha = 0.01$，$\beta = 0.95$ の場合，52 例必要となる．

証明力の強さから Evidence のタイプは表 4 のように分類される．

表 3: Evidence のタイプ

Ia: ランダム化比較試験のメタ分析による
Ib: 少なくとも 1 つのランダム化比較試験による
Ic: 対象者がすべて死亡または死亡者なしの場合
IIa: 少なくとも 1 つのよくデザイン化された
非ランダム化比較試験のメタ分析による
　b: 少なくとも 1 つのよくデザイン化された
非ランダム化比較試験による
　c: 少なくとも 1 つの他のよくデザイン化された
準実験的研究による
　a: 比較研究や相関研究，症例対照研究等よく
デザインされた非実験的記述研究のメタ分析
　b: 比較研究や相関研究，症例対照研究等よく
デザインされた非実験的記述研究
　: 症例研究/質の低いコホート，症例対照研究
　: 専門家委員会の報告や意見，あるいは
権威者の臨床試験

メタ分析は，系統的レビューの中で，そのレビューした論文の検定の結果を統合して，一般的な検定の結果を導きだそうというものである．系統的レビューはある問題に対して，これまでなされた研究成果を集積し，全体として結論を導くアプローチであるが，一般的なレビューと以下の点で異なる:

1. 一般的なレビューと異なり，仮説を 1 つに絞る
2. どのデータベース，検索方法を用いたか詳細に記載する．
3. 論文の選択基準を明確にする．
4. とりあげた論文は批判的に評価する．
5. 結果を量的に要約する．

6. 解釈・評価は Evidence に基づく

最後の 2 項における解釈・評価の段階でメタ分析を行うが，メタ分析は，ある証明力をもつ研究デザインを集めて，結果を統合することによって Evidence としての汎化能力を高めるものと考えることができる．

EBM とアクティブマイニング

前節までで論じたように，EBM は体系的に観察・収集されたデータに基づく医療を指向したものであるが，EBM の効率的な実現にはデータをアクティブに収集，解析，解釈するプロセスの高度化が不可欠であり，この目標はまさしく，アクティブマイニングの実現に一致するものである．

表 4 は，これまでの議論をまとめて，アクティブマイニングと EBM との対応関係とを示したものてある．必ずしも EBM はらせんモデルとして提起されているものではないが，表の対応づけによって，EBM がアクティブマイニングの医療応用として位置づけることが可能である．この対応付けに基づいて，EBM の実践手順とアクティブマイニングのプロセスを重ね合わせたものが，図 2 であり，アクティブマイニングの医療応用は，図示されたプロセスのコンピュータへ実装による EBM の効率化実現を目標としたものである．

表 4: アクティブマイニングと EBM との対応関係

	EBM	アクティブマイニング
Step1:	疑問の定式化	問題設定 (アクティブユーザーリアクション)
Step2:	Evidence の収集	情報源からのアクティブ情報収集 ユーザ指向アクティブマイニング
Step3:	Evidence の質の評価 (Evidence の「客観的」評価)	ユーザ指向アクティブマイニング
Step4:	適用性判断 (Evidence の「主観的」評価)	アクティブユーザーリアクション

おわりに

アクティブマイニングと EBM との対応関係について概説した．アクティブマイニングは一般的な枠組みとして，データ解析およびそれに基づく知識の発見の大きな枠組みを構想しているが，EBM はその医療応用として実現可能である．EBM の効率的な実現は，21 世紀の日本の医療に大きな役割を果たすことが予想され，アクティブマイニングは医療の IT 化の一端を担う重要な情報技術となるであろう．

参考文献

[1] 稲田裕, 野崎貞彦. 新簡明衛生公衆衛生. 南山堂 (1994).

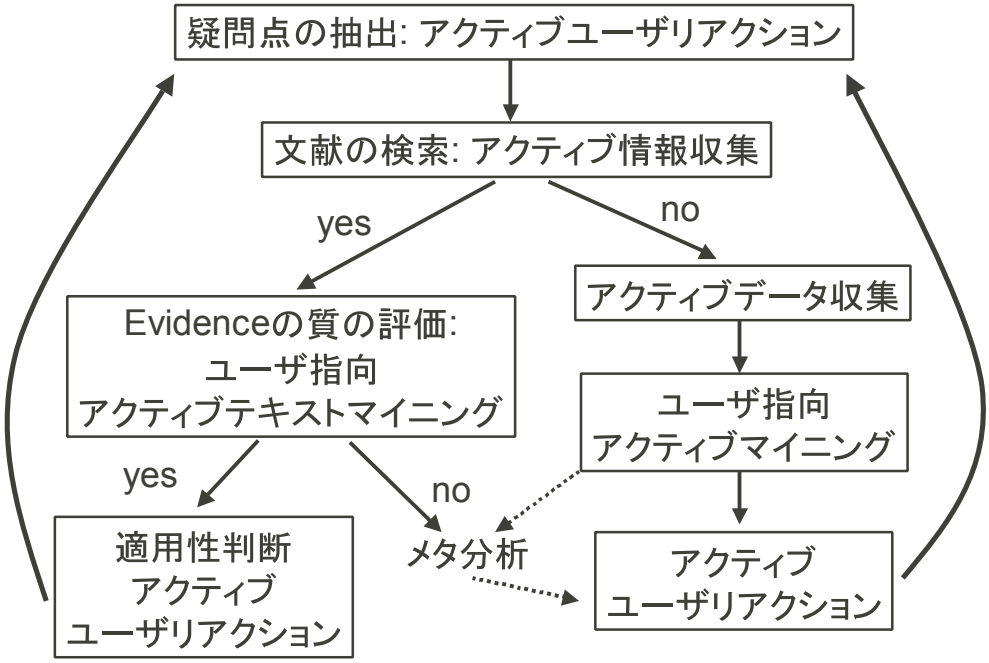


図 2: アクティブマイニングによる EBM の実現

[2] 福井次矢: EBM 実践ガイド , 医学書院 (2000).

[3] Petitti, D.B. *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis*, Oxford University Press, Oxford (1994).

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

研究代表者　津本 周作　(島根医科大学医学部医療情報学講座)
研究分担者　高林 克日己　(千葉大学医学部附属病院医療情報部)
柳樂 真佐実　(島根医科大学医学部医療情報学講座)
平野 章二　(島根医科大学医学部医療情報学講座)

背景と目的

近年，各種臨床検査機器のデジタル化とネットワーク化が急速に進展し，血液検査データ，生化学検査データ，画像検査データなど様々な検査データを自動的に収集・データベース化することが可能となった．多数の患者を抱える拠点診療施設では膨大な検査データが既に蓄積されており，多様な症例を含む大規模検査データベースが構築されている．しかし，このように蓄積されたデータは，過去の検査履歴を参照する目的で個別に利用されることが多く，サンプル数の多さから生まれる様々な利点を十分に活用できていない現状にある．このため，多数の症例を効率的かつ多面的に比較検証する技術の確立と，得られる有益な診療知識をより質の高い医療に結びつけるための方策づくりが急務となっている．このような背景の中，大規模データベースからの知識獲得を主眼に置くデータマイニング技術は一層その重要性を増しており，EBM （Evidence-based Medicine，根拠に基づく医療）を支援する有力な手段として精力的に研究が進められている [1]-[15].

一方，このようなデータベース化の流れが定着するにつれ，同一患者の検査データを数年から数十年の長期にわたり継続的に収集した時系列検査データも利用可能になりつつある．このような時系列検査データは，数日を単位とする短期間の推移のみならず，年単位の長期にわたる検査値推移パターンと疾患との対応関係を示すものであるため，その解析により，慢性疾患を誘発する要因の特定，あるいは発病時期の予測等が可能になると期待される．しかしながら，これらのデータは当初から解析を目的に収集されたものではなく，以下の要因を含む不均質なものであるため，現在のところ有効には活用されていない：

- 検査の有無による欠損値の存在
  全ての受診日に検査を行う訳ではなく，検査しない日も存在する．また，検査項目も必要性に応じて変化する．
- 不定期な収集間隔
  症状の増悪・軽快の状況と患者，病院双方の時間的都合により，受診間隔が数日〜数ヵ月まで不規則に変化する．
- ノイズ
  検査当日の僅かな体調変化が検査値を変化させる．

我々は，これらの要因を含む不均質な時系列臨床検査データの解析法として，多重スケールマッチング [16] とラフクラスタリング [17, 18, 19, 20] を組み合わせた新たな解析

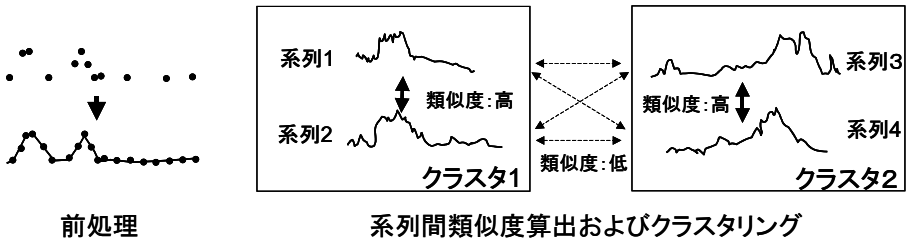


図 1: 本方法の概略

法を開発している [21, 22, 23, 24]．多重スケールマッチングは対象間の類似性を様々な視野スケールで部分ごとに比較する方法であり，時系列データに適用した場合，検査値の推移パターンの類似性を長期的，短期的両方の観点から調べることができる．マッチングは隣接する変曲点を両端とする部分系列であるセグメントを単位として行われるため，検査値が上昇していた期間，下降していた期間を基準に類似性を調べることが可能となる．

一方，ラフクラスタリングはラフ集合論 [25] の識別不能性に基づくクラスタリング法であり，相対的類似度のみで類似性が定義されるデータにおいても可読性の高いクラスタを生成することができる．本方法では，これらを組み合わせ，多重スケールマッチングにより得られる系列間の類似性を基にラフクラスタリングを適用し，全系列をいくつかの代表的な変化パターンに分類する．その後，得られたパターンと疾患の組合わせを比較することで，特定の疾患と関連を持つ検査値の推移パターンを発見する．

本稿ではまず，方法の概略に続いて述べた後，時系列臨床検査データの説明，前処理，多重スケールマッチングとラフクラスタリングの各処理についてそれぞれ順に述べる．続いて実データにおける実験結果を示し，最後にこれまでに得られた結果と今後の展望をまとめる．

解析方法

全体の流れ

本方法の概略を図 1 に示す．まず，前処理としてデータのリサンプリングを行い，等間隔でサンプリングした新たなデータを構築する．これにより，欠損値を補完するとともに，検査値の上昇/下降の周期を同一尺度で表現し，その大きさによる比較を可能とする．次に，任意の検査項目について，患者間における検査値推移パターンの類似性を多重スケールマッチングにより求める．全ての患者組について系列間類似度を求めた後，得られた類似性を基準としてラフクラスタリングを適用し，類似した系列をまとめて全体を幾つかのクラスタに分類する．このようにして得られるクラスタと診断クラスとを比較することで，疾患と関係する特徴的な検査値の推移パターンを可視化する．

時系列臨床検査データの概略と前処理

表 1 に時系列臨床検査データの例を示す．データの各行は各受診日に対応し，患者 ID，受診日とともに当日の検査データがそれぞれ記録されている．受診日は数日から数ヵ月の範囲で不規則に変化しており，５月６日，７日に見られるように検査の実施されなかった

表 1: 時系列検査データの例

PATID	Date	GOT	GPT	LDH	ALP	TP	ALB	UA	UN
0001	860419								
0001	860430	25	12	162	76	7.9	4.6	4.7	18
0001	860502	22	8	144	68	7	4.2	5	18
0001	860506								
0001	860507								
0001	860508	22	13	156	66	7.6	4.6	4.4	15
0001	860512	22	9	167	64	8	4.8	4.5	14
0001	860519	28	13	185	60	7.5	4.5	4	13
0001	860526	21	12	134	56	7.2	4.4	3.7	16
0001	860527	23	10	165	55	7.1	4.2	3.6	14
0001	860528								
0001	860630	23	10	137	66	7.6	4.4	3.2	12

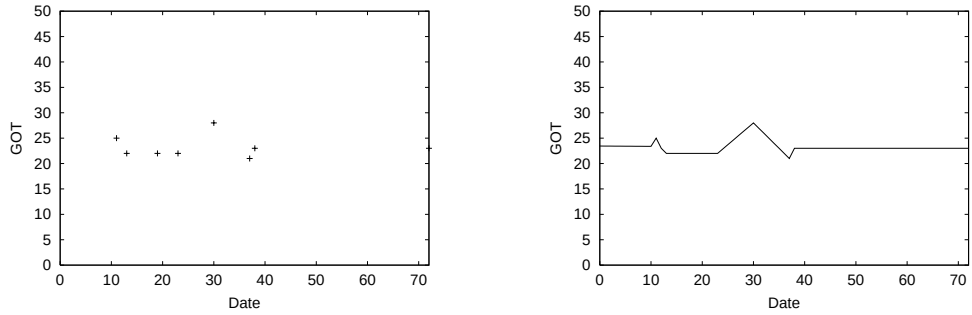


図 2: 補間前後の GOT データ．左：補完前　右：補完後

日も存在する．また，受診日，検査項目等は患者の時間的制約および病態に応じて変化する．

本方法では，異なる患者の検査値推移パターンを比較するために，同一のサンプリング間隔を用いて系列をリサンプリングする．サンプリング間隔は対象とする疾患の特性に基づき決定し，急性疾患の場合は数日，慢性疾患の場合は数ヶ月程度となる．一般に，受診日が連続しているケースは稀であり，また検査項目にもばらつきがあるため，元データには欠損値が含まれる．このため，リサンプリング時には補完処理が必須となる．時系列データの補間には，系列平均，周囲値平均，線形補間，トレンド，自己相関等 [26] が提案されているが，ここでは臨床検査データに特徴的に見られるデータ収集間隔の不定期性を考慮し，線形補間とトレンドを組み合わせた方法を用いる．

1. 病態から急な検査が不要と判断されている場合：
   この場合，何らかの異常が疑われて初めて必要な検査が実施されることとなる．そこで，最初に検査が行われるまでの期間の欠損値は，大きな変化を伴わず，全体の傾向を反映し滑らかな時間推移をたどるとみなし，系列全体の線形トレンドで補間する．

2. 前検査との間隔が近く，次回以降の受診日に検査を延期している場合：
   この場合，検査値は前後の検査日のものと関連して変化するとみなし，近傍有効値の線形補間により補間する．

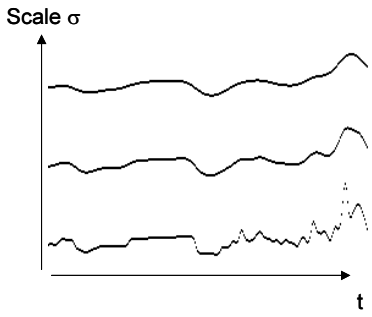


図 3: 多重スケール表現

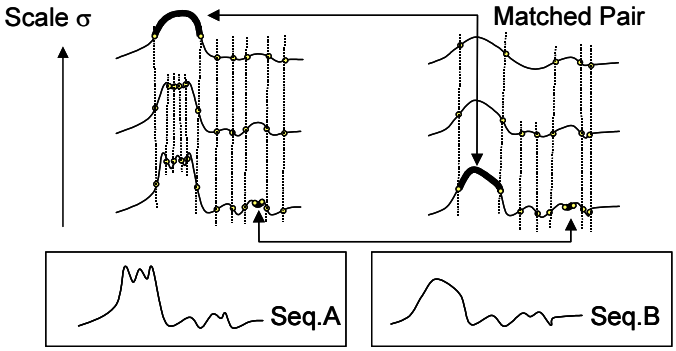


図 4: 多重スケールマッチング

3. 治療後の経過観察のため，間隔を開けて検査している場合：
この場合も 1. と同様に，全体的な検査値に大きな変化はないとみなし，系列全体の線形トレンドを用いて補間する．

表 1 の GOT データプロットおよびその補間結果を図 2 にそれぞれ示す．この例では，2 週間以内に有効な検査データが存在する場合は近傍値による線形補間，それ以外の場合には線形トレンドを用いて補間している．また，横軸の日付は初回検査日である 4 月 19 日からの相対日数であり，0-10 日目までは上記 1，11-37 日目までが上記 2，38 日目以降が 3 の期間にそれぞれ該当する．

多重スケールマッチング

Mokhtarian [27] らにより提案された多重スケールマッチングは，対象図形を様々な視野スケールで記述，比較する方法である．マッチングは部分輪郭ごとの類似性を基準にして行われ，対応する部分輪郭組は同一スケールのみならず，異なるスケールに渡って探索される．これにより，局所的な類似性だけでなく，より大局的な観点から観察した類似性に基づきマッチングを行うことが可能となる．この方法ではスケールを連続的に変化させる必要があり，計算量の問題が指摘されていたが，上田ら [16] が変曲点間の凹凸セグメントをマッチング単位とすることで離散スケールの導入を可能とし，この問題を解決した．本方法では，上田らの方法を用いて患者間での検査値系列のマッチングを行う．ここでは，検査値の増減に起因して生じる系列の凹凸構造を部分輪郭の凹凸構造と対応させる．これにより，短期的な変化パターンの類似性のみならず，より長期的な変化パターンの類似性を評価する．

まず，時刻 t をパラメータとする関数 $x(t)$ で検査値の系列を表現する．このとき，スケール σ における系列は，$x(t)$ とスケールファクター σ をもつガウス関数 $g(t,\sigma)$ との畳み込みとして以下のように定義される．

$$\begin{aligned} X(t,\sigma) &= x(t) \otimes g(t,\sigma) \\ &= \int_{-\infty}^{+\infty} x(u) \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-u)^2/2\sigma^2} du. \end{aligned}$$

図 3 に σ を変化させた場合の系列の変化を示す．同図および上式から明らかなように，スケールの増加とともに近傍値との平滑化が進み，より変曲点の少ない滑らかな系列が得られる．系列上の各点における曲率は次式で与えられる．

$$K(t,\sigma) = \frac{X''}{(1+X'^2)^{3/2}},$$

ここで，X'，X'' は $X(t,\sigma)$ の t による 1 次および 2 次微分である．$X(t,\sigma)$ の m 次微分 $X^{(m)}(t,\sigma)$ は，$x(t)$ と $g(t,\sigma)$ の m 次微分 $g^{(m)}(t,\sigma)$ の畳み込みとして次式により与えられる．

$$X^{(m)}(t,\sigma) = \frac{\partial^m X(t,\sigma)}{\partial t^m} = x(t) \otimes g^{(m)}(t,\sigma).$$

次に，曲率の符合の変化から系列上の変曲点の位置を求め，隣接する変曲点を両端とする凹凸セグメントを構築する．スケール $\sigma^{(k)}$ における検査値系列 $\mathbf{A}^{(k)}$ を N 個のセグメントの集合とすると，

$$\mathbf{A}^{(k)} = \left\{ a_i^{(k)} \mid i = 1, 2, \cdots, N^{(k)} \right\}.$$

ここで，a_i^k はスケール $\sigma^{(k)}$ における i 番目のセグメントを示す．同様に，スケール $\sigma^{(h)}$ における比較対象系列を $\mathbf{B}^{(h)}$ とすると，

$$\mathbf{B}^{(h)} = \left\{ b_j^{(h)} \mid j = 1, 2, \cdots, M^{(h)} \right\}$$

と表現できる．このとき，セグメント $a_i^{(k)}$ と $b_i^{(h)}$ の相違度 $d(a_i^{(k)}, b_j^{(h)})$ を次式により定義する．

$$d(a_i^{(k)}, b_j^{(h)}) = \frac{\mid \theta_{a_i}^{(k)} - \theta_{b_j}^{(h)} \mid}{\theta_{a_i}^{(k)} + \theta_{b_j}^{(h)}} \left| \frac{l_{a_i}^{(k)}}{L_A^{(k)}} - \frac{l_{b_j}^{(h)}}{L_B^{(h)}} \right|,$$

ここで，$\theta_{a_i}^{(k)}$ および $\theta_{b_j}^{(h)}$ は各セグメントに沿った接ベクトルの回転角，$l_{a_i}^{(k)}$ および $l_{b_j}^{(h)}$ は各セグメントの長さ，$L_A^{(k)}$ および $L_B^{(h)}$ は対象系列 A，B のスケール $\sigma^{(k)}$，$\sigma^{(h)}$ における総セグメント長をそれぞれ示す．すなわち，セグメント回転角と長さの寄与の差が大きいほど高い相違度が与えられる．複数のセグメントが置換されてできたセグメント同士の相違度も同様に定義される．本方法では，マッチング終了後の残相違度の逆数を系列間類似度とした．

多重スケールマッチングにおけるマッチング手続きは，全てのセグメント組から相違度の総和を最小にする組を探索することに相当する．図 4 上側に示すマッチング例では，系列 A の 5 つの連続したセグメントが上位スケールで 1 つのセグメントに置換され，これが系列 B の 1 セグメントと対応している．一方，同図下側に示すもう 1 つのマッチング例では，最下位スケールおいて対応するセグメントが見られる．このように，短期的に類似した傾向が見られる場合は下位スケールで，短期的には異なるが長期的には類似した傾向が見られる場合はより上位のスケールで対応がとられる．なお，マッチングアルゴリズムの詳細については文献 [16] を参照されたい．

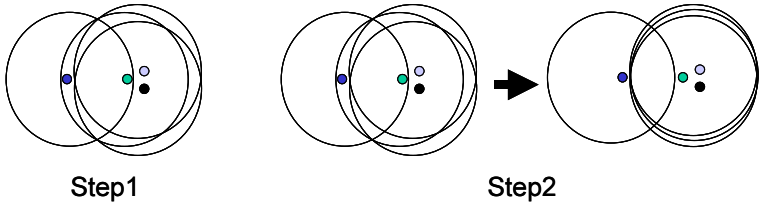


図 5: ラフクラスタリングの流れ

ラフクラスタリング

対象間の類似度が原点を持たない相対的類似度として与えられる場合，クラスタ内の分散，重心等を定義することが困難で，クラスタとしてのまとまりを評価することは容易ではない．ラフクラスタリングは，ラフ集合論の識別不能性の概念に基づくクラスタリング法であり，対象のまとまり具合いを識別不能度として表現することで，相対的類似度で表現されたデータにおいても可読性の高いクラスタを生成することができる．多重スケールマッチングにより得られる類似度（相違度）は，任意の 2 検査系列間の類似性を表す相対的な尺度であるため，本方法ではラフクラスタリングを適用して系列を分類する．

ラフクラスタリングは，(1) 初期同値関係の構築，(2) 同値関係の再帰的更新，の 2 ステップから構成される．図 5 に各ステップの概略を示す．第 1 ステップでは，各対象に対して自らと類似したものと異なるものを分類する初期同値関係を与える．n 個の対象からなる全体集合を $U = \{x_1, x_2, ..., x_n\}$ としたとき，対象 x_i に対する同値関係 R_i は次式により定義される．

$$R_i = \{\{P_i\}, \{U - P_i\}\},$$

$$P_i = \{x_j |\ s(x_i, x_j) \geq S_i\}, \quad \forall x_j \in U.$$

ここで，P_i は x_i と類似した対象の集合であり，類似度が閾値 S_i を越える対象の集合として定義される．類似度 s は対象間の類似度で，その閾値 S_i は類似度が著減する位置に決定される．クラスタは，全ての同値関係を用いても識別不能な対象の集合（同値類）として定義される．図 5 はこれらをユークリッド空間上で表現したもので，各対象を中心とする円の半径は S_i に相当し，この中に存在する他の対象はすべて同値類とみなされる．2 つの対象間を交差する円がただ 1 つも存在しない場合，その対象は同一クラスタに分類される．同図の例の場合，4 つの対象が 3 つのクラスタに分類されている．なお，時系列データの解析では，各オブジェクトが各系列に対応し，類似度 s として前節で述べた多重スケールマッチングの結果得られる 2 系列間の類似度を用いる．

第 2 ステップでは，第 1 ステップで個々に構築した初期同値関係を全体的な観点から修正し，より可読性の高いクラスタを生成する．まず，ある 2 つの対象 x_i と x_j が，他のどれだけ多くの対象から識別不能と見なされているかを示す識別不能度 γ を次式により定義する．

$$\gamma(x_i, x_j) = \frac{1}{|U|} \sum_{k=1}^{|U|} \delta_k(x_i, x_j),$$

$$\delta_k(x_i, x_j) = \begin{cases} 1, & \text{if } [x_k]_{R_k} \cap ([x_i]_{R_k} \cap [x_j]_{R_k}) \neq \phi \\ 0, & \text{otherwise.} \end{cases}$$

ここで，$[x_i]_{R_i}$ は同値関係 R_i において x_i と同値類とみなされる対象の集合を示す．識別不能度 γ が高い対象は類似度が高く，同一のクラスタに分類されることが望ましい．逆に，識別不能度の高い対象を異なるクラスタに類別するような同値関係 R_i は詳細すぎる類別知識を与えているといえる．そこで，そのような同値関係を以下の手続きにより R_i' に修正し，詳細すぎる類別知識による細かなクラスタの生成を抑制する．

$$R_i' = \{\{P_i'\}, \{U - P_i'\}\}$$

$$P_i' = \{x_j | \gamma(x_i, x_j) \geq T_h\}, \quad \forall x_j \in U.$$

ここで，T_h は対象を識別不能と見なす閾値であり，類別知識の粗さに対応づけられる．この T_h の値を徐々に低下させつつ再帰的に同値関係を更新することで，適度に粗い知識に基づくクラスタリング結果が得られる．図 5 の例では，2 つの同値関係を更新することでクラスタ数が 4 から 2 に減少している．なお，2 度目以降の同値関係の更新は，初期同値関係ではなく前回更新後の同値関係を用いる．

結果

今回は初期実験として膠原病データベース [28] から得られた時系列 GOT 検査データに本方法を適用し，多重スケールマッチングの時系列データ解析への適用可能性について調べた．簡単のため，データセットには，比較的特徴的な変化がみられる 10 程度の系列を視覚的に選択して用いた．図 6 に同一クラスタに分類された系列組の 1 例を示す．同図において，上段，下段の s1-s3 はスケール $\sigma = 1.5, 4.5, 7.5$ における系列をそれぞれ表し，result はマッチング結果を示す．セグメント A1 はセグメント B1 と対応し，同様に，セグメント An はセグメント Bn と対応している．また，横軸は時間軸であり，データのリサンプリング間隔は 1 日とした．

同図から，本方法が系列間に見られる推移パターンの類似性を正しく認識できていることがわかる．例えば，セグメント A3 および B3 から始まる一連の部分系列は，小幅な上昇 (A4,B4)，中程度の下降 (A5,B5)，大幅な上昇 (A6,B6)，小幅な下降 (A7, B7) など，類似した推移パターンをとっている．また，(A6,B6) に見られるように，原系列のサンプリング間隔の違いに起因する局所的相違が上位のスケールにおいて吸収され，より大局的な特徴を認識できていることがわかる．

図 7 に別クラスタに分類された系列を示す．このように異なる収集期間をもつ系列に対しても，推移パターンの特徴認識が可能であった．これらから，多重スケールマッチングを時系列データ解析に用いることで，一つの視野スケールのみではなく短期的，長期的な特徴を同時に考慮した系列比較が可能となることが示された．

むすび

本稿では，多重スケールマッチングとラフクラスタリングの組合わせによる時系列臨床検査データベースの解析法を提案した．検査値の変化には，ごく短期の内に生じるものと

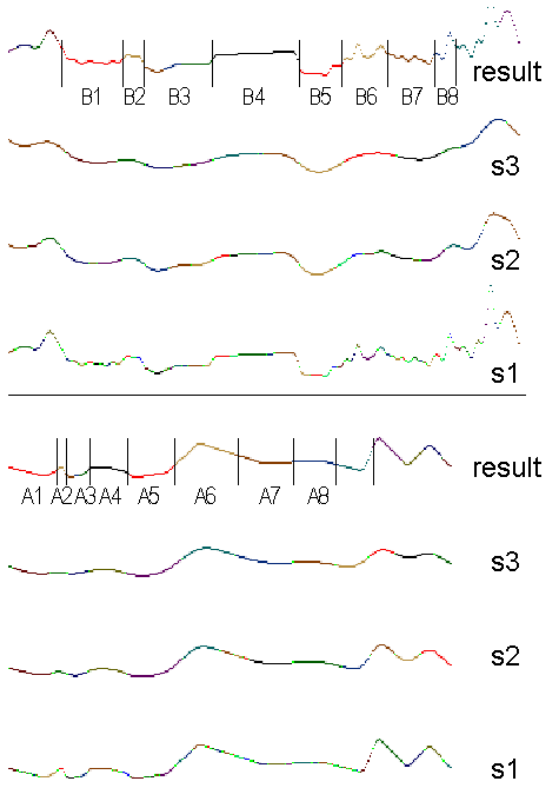


図 6: GOT データに対する適用結果 (クラスタ 1)

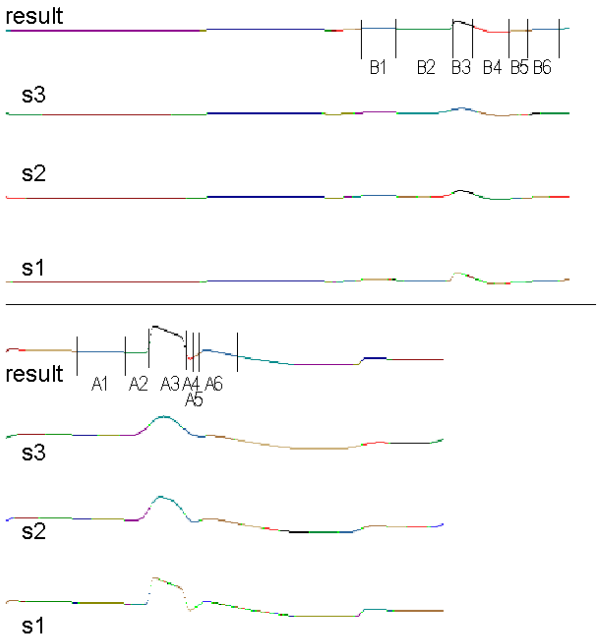


図 7: GOT データに対する適用結果 (クラスタ 2)

長期にわたるものがあるため，それぞれの期間を考慮した解析法が必要となる．多重スケールマッチングは，このような要求を満たす方法一つであり，様々な時間スケールにおける比較を効率的に実現することができる．一方で，系列間に絶対的な原点と方向をもつ類似度（あるいは距離）を定義することは困難であるため，相対的な類似度を用いてこれらを分類する方策も必要である．識別不能性を基礎概念とするラフクラスタリングは，このような場合にも効果的に機能する方法の一つであると言える．本方法は，これらの特徴を組み合わせたものであり，時系列臨床検査データの有効な解析手段となりうるものである．今後，多様なデータベースに適用し，その有効性を検証していくとともに，属性選択法 [29, 30] についても検討を進める予定である．

参考文献

[1] L. Polkowski, S. Tsumoto, and T.Y. Lin (eds.): Rough Set Methods and Applications : New Developments in Knowledge Discovery in Information Systems, Physica-Verlag, New York (2001).

[2] S. Tsumoto: Discovery of Clinical Knowledge in Databases Extracted from Hospital Information Systems, K.J. Cios (ed.) Medical Data Mining and Knowledge Discovery,

Physica-Verlag, New York, pp. 433–454 (2001).

[3] S. Tsumoto: Induction of Rule about Complications with the Use of Rough Sets, W. Pedrycz (ed.) Granular Computing: an emerging paradigm, Physica-Verlag, New York, pp. 384–397 (2001).

[4] S. Tsumoto: Chapter G5: Data Mining in Medicine, W. Kloesgen and J. Zytkow (eds.) Handbook of Data Mining and Knowledge Discovery, Oxford University Press (2001).

[5] T. Terano, T. Nishida, A. Namatame, S. Tsumoto, Y. Ohsawa, and T. Washio (eds.): New Frontiers in Artificial Intelligence, Joint JSAI 2001 Workshop Post-Proceedings, Lecture Notes in Computer Science 2253, Springer-Verlag, Heidelberg (2001).

[6] S. Yokoyama, K. Matsuoka, S. Tsumoto, M. Harao, T. Yamakawa, K. Sugahara, C. Nakahama, S. Ichiyama, and K. Watanabe: Study on the Association between the Patient's Clinical Background and the Anaerobes by Data Mining in Infectious Disease Database, BMFSA, vol. 7, no. 1, pp. 121–128 (2001).

[7] 安田 晃, 柳樂真佐実, 孫 暁光, 津本周作, 山本和子: 自主学習における学生の自己評価の変動に関する解析, 医学教育, vol. 32, pp. 69–75 (2001).

[8] 津本周作: 医学における知識発見手法の可能性 (特集: データマイニングコンテスト), 情報処理, vol. 42, pp. 472–477 (2001).

[9] 津本周作: ラフ集合論の現状と課題 (特集: ラフ集合の理論と応用), 日本ファジィ学会誌, vol. 13, pp. 552–561 (2001).

[10] 鈴木英之進, 菅谷信介, 津本周作: サポートベクターマシンに基づく医療データからの事例発見, オペレーションズ・リサーチ, vol. 46, no. 5, pp. 243–248 (2001).

[11] S. Tsumoto: Medical Knowledge Discovery in Hospital Information System, Proceedings of SPIE vol.4384, Data Mining and Knowledge Discovery: Theory, Tools, and Technology III, pp. 229–237 (2001).

[12] S. Tsumoto: Statistical Extension of Rough Set Rule Induction, Proceedings of SPIE vol.4384, Data Mining and Knowledge Discovery: Theory, Tools, and Technology III, pp. 180–188 (2001).

[13] S. Tsumoto: Mining Positive and Negative Knowledge in Clinical Databases Based on Rough Set Model, Proceedings of the fifth European Conference on Principles of Knowledge Discovery in Databases(PKDD2001), pp. 460–471 (2001).

[14] 津本周作: 医療情報から見た歯科医療界の今後-最新の情報技術と診療支援-, 島根県歯科医師会生涯教育講座, pp. 1–12 (2001).

[15] 津本周作: 特別講演: 医療におけるアクティブマイニング, –Medical Data Mining からの新たな展開–, バイオメディカルファジイシステム学会第 14 回年次大会講演論文集, pp. 1–4 (2001).

[16] 上田修功, 鈴木 智: 多重スケールの凹凸構造を用いた変形図形のマッチングアルゴリズム, 電子情報通信学会論文誌 (D-II), vol. J73-D-II, no. 7, pp. 992–1000 (1990).

[17] S. Hirano, S. Tsumoto, T. Okuzaki, and Y. Hata: A Clustering Method Based on Rough Set Theory and Its Application to Knowledge Discovery in the Medical Database, MEDINFO, vol. 10, pp. 206–210 (2001).

[18] S. Hirano and S. Tsumoto: A Knowledge-Oriented Clustering Technique Based on Rough Sets, Proceedings of the 25th IEEE International Computer Software and Applications Conference (Compsac2001), pp. 632–637 (2001).

[19] S. Hirano and S. Tsumoto: Indiscernability Degrees of Objects for Evaluating Simplicity of Knowledge in the Clustering Procedure, Proceedings of the 2001 IEEE International Conference on Data Mining, pp. 211–217 (2001).

[20] 平野章二, 津本周作: ラフ集合論に基づく知識指向型クラスタリング法, バイオメディカルファジイシステム学会第 14 回年次大会講演論文集, pp. 6–9 (2001).

[21] S. Tsumoto: Temporal Knowledge Discovery in Time-Series Medical Databases based on Fuzzy and Rough Reasoning, Proceedings of Ninth International Fuzzy Systems Association World Congress(IFSA'01), (CD-ROM) (2001).

[22] S. Tsumoto: Discovery of Temporal Knowledge in Medical Time-Series Databases Using Moving Average, Multiscale Matching, and Rule Induction, Proceedings of the fifth European Conference on Principles of Knowledge Discovery in Databases (PKDD2001), pp. 448–459 (2001).

[23] S. Hirano and S. Tsumoto: Analysis of Time-series Medical Databases Using Multiscale Structure Matching and Rough Sets-based Clustering Technique, Proceedings of the 2001 IEEE International Conference on Fuzzy Systems (2001).

[24] 平野章二, 津本周作: 多重スケールマッチングとラフクラスタリングによる時系列臨床検査データベースの解析, 人工知能学会第 46 回人工知能基礎論研究会, 第 54 回知識ベースシステム研究会合同研究会資料 (SIG-FAI/KBS-J-42), pp. 257–260 (2001).

[25] Z. Pawlak: Rough Sets, Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht (1991).

[26] R.H. Shumway and D.S. Stoffer: Time Series Analysis and Its Applications, Springer-Verlag, New York (2000).

[27] F. Mokhtarian and A. K. Mackworth: Scale-based Description and Recognition of planar Curves and Two Dimensional Shapes, IEEE Transactions on Pattern Analysis Machine Intelligence, vol. PAMI-8, no. 1, pp. 24–43 (1986).

[28] URL: http://www.shimane-med.ac.jp/med_info/open_data/

[29] S. Tsumoto, S. Hirano, A. Yasuda, and K. Tsumoto: Analysis of Amino Acid Sequences by Statistical Technique, Proceedings of the 4th Conference on Computational Biology and Genome Informatics (CBGI-02), 2002 (in press).

[30] 孫 暁光, 柳樂真佐実, 平野章二, 安田 晃, 津本周作: 臨床データベース解析のための類似性尺度とその評価, 第 21 回医療情報学連合大会講演論文集, pp. 504–505 (2001).

カスケードモデルの発展と
発ガン性・変異原性を示す分子の発見

研究代表者　岡田孝（関西学院大学情報メディア教育センター）

背景と目的

化学物質が示す生理活性をその分子構造から予測することが，基本的な課題である．例えば，一種の化合物に対するネズミの発ガン性をチェックするだけで数億円の費用がかかる．１０万種程度の化学物質がかなりの量で生産されており，これら物質すべての発ガン性を生物実験によって調査することは不可能である.そこで，化合物構造から人体や他の動植物に対する生理活性を予測することが求められている．薬品類についても同様である．すなわち，よく使われる薬品でも，特定の疾患を有する人には毒となる．反対に，ありふれた化学物質にこれまで知られていない薬効が見つかる場合もある．

本研究では，化学物質群の構造・性質とその活性のデータベースから，（１）それぞれの生理活性に対して特徴的な部分構造や物理化学的性質を見出し，（２）新規の化学物質群における未知の生理活性を予測して，危険を回避することを目的とする．

本年度の研究は以下の２項目について実施した．項目ごとに，その検討内容および結果と考察を述べる．

(1) 理解容易なルール群表現を目指したカスケードモデルの発展
(2) 変異原性および発ガン性に関連する化合物特徴の認識

１．理解容易なルール群表現を目指したカスケードモデルの発展

検討内容

(1) どのようなルール表現がわかりやすいか

相関ルールによるマイニングでは，多様な視点からのデータ特徴を抽出し，それをルールとして表現することができる [1]．反面，実際に適用した場合は，非常に多数のルールが出力され，利用者にとって解釈が困難になることも周知の事実である．

相関ルールを発展させたカスケードモデルでは，支持度と確信度に代わって，ルール強度を *BSS* という一つの数値で表せるため，ルール群の解釈が基本的に容易となる [2,3]．しかし，利用者サイドの立場からすれば，やはり多くのルールが互いに独立してデータの特徴点を指摘するため，データの全貌を把握する立場からは問題がある．

多変量解析では，変数間に相関の高いものが存在するときに数値計算面，解釈面において各種の問題が起こることが，共線性の問題としてよく知られている．この点は決定木においても同じである．相関ルールやカスケードモデルにおいては，各変数を独立に扱うため共線性自体は問題とならない．反面，同一事象を複数のルールが表現することになる．このように考えると，ルール数の多さというマイニングの課題は，共線性が形を変えて現れたものとも言える．

著者は，データの全体像を datascape と名付け，これを利用者が容易に見渡し，その

特徴点を把握できるようなシステムの構築を目指している．アクティブマイニングの標的の一つに，「発掘された知識をユーザにとって素早くかつ容易に把握可能な形で表示し，対象への理解を深め，知的創造性を刺激し，それにより新たな知識発掘の糸口を誘発する．」がある．この目的の達成には，datascape を与えるシステムが必要となろう．

本節では，この datascape を見渡せるようにするため，カスケードモデルにもとづくルール群表現にいくつかの新機能導入を提案する．以下にカスケードモデルおよびそのルール群表現法を簡単に説明した後，新たに提案するルール群表現を解説する．最後に，この方法により髄膜炎データセットを解析した結果を例として提示する．

(2) カスケードモデル

<u>モデルの枠組み</u>

このモデルは，相関ルールと同様に itemset のラティスを構築し，ラティス内で識別力の高いリンクをルールとして選択し提示する．その際，item としてはすべて [attribute: value] の形式を採用する．このモデルの第 1 の特徴は，ラティス内の節点に potential を，リンクに power を定義することにある．Gini によるカテゴリー変数に対する平方和（SS）の定義式 (1) から導かれる分散を potential と考え，(2) 式による分解で現れる (3) 式の *BSS*（Between-groups sum of squares）をリンクの power としている [5, 6].

$$SS_i = \frac{n}{2}\left(1 - \sum_a p_i(a)^2\right) \quad , \tag{1}$$

$$TSS_i = \sum_g \left(WSS_i^g + BSS_i^g\right) \quad , \tag{2}$$

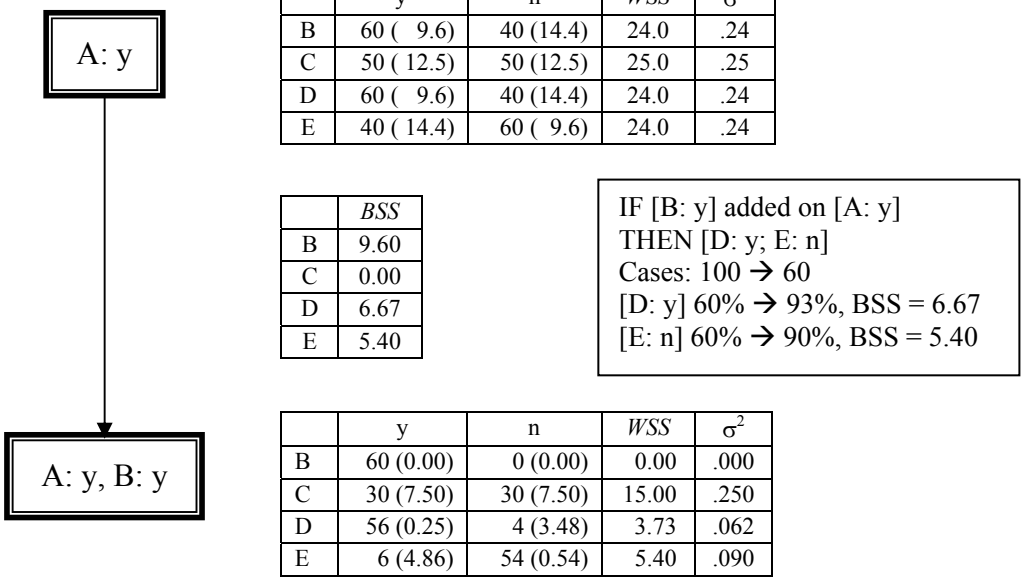
$$BSS_i^g = \frac{n^L}{2} \sum_a \left(p_i^L(a) - p_i^U(a)\right)^2 \quad . \tag{3}$$

ここで，$p_i(a)$は属性 i の値が a である確率を示し，g がグループを，U はグループに分割前の節点，L は分割後の節点を表す．

属性 A-E からなる問題で，図 1.1 のように[A: y]のアイテムを持つ上節点と[A: y, B: y]を持つ下節点，およびその間のリンクを考えよう．節点の itemset に現れない item を veiled item と呼ぶ．それぞれの itemset をサポートする instance 中の各 veiled item を数えれば，３種の表に示すように，節点の *WSS*, リンクの *BSS* を評価することができる．大きな *BSS* 値に対応する属性がリンクに沿って付加された item: [B: y]と大きな相互作用を持つから，これをルールとして表現すればよい．

この例におけるルール表現は，図中右側の四角内のように与えられる．ここで，LHS 部は下節点上の itemset に含まれる item を表し，その中でもルールリンクにおいて付加された item [B: y] が主条件として強調されている．他方，RHS 部の表現は，解析目的にそった属性中から *BSS* 値の大きなものを選択して表示する．

ところで，ラティスを基盤としたシステムでは，常に節点数の組み合わせ爆発が問題となる．しかし，前節で述べた *BSS* 値を *minimum support* 条件と併せて用いることにより，効率的な枝狩りをできることがすでに示されている [7].



	y	n	*WSS*	σ^2
B	60 (9.6)	40 (14.4)	24.0	.24
C	50 (12.5)	50 (12.5)	25.0	.25
D	60 (9.6)	40 (14.4)	24.0	.24
E	40 (14.4)	60 (9.6)	24.0	.24

	BSS
B	9.60
C	0.00
D	6.67
E	5.40

	y	n	*WSS*	σ^2
B	60 (0.00)	0 (0.00)	0.00	.000
C	30 (7.50)	30 (7.50)	15.00	.250
D	56 (0.25)	4 (3.48)	3.73	.062
E	6 (4.86)	54 (0.54)	5.40	.090

図１．カスケードモデルにおけるリンクとルール

ルール表現と問題点

カスケードモデルでは，利用者の指定値以上の *BSS* を持つリンク群から，解析に有効と考えられるものをルールとして表示する．ルール群は複数のルールセットに分けて表示される．ここで，一つのルールセットは，その中のルール群がデータベース全体の事例をできる限り広くカバーするように選択される．

ルールセット選択のアルゴリズムを Algorithm 1 に示す．ここで，最初に与えられる links が大きい *BSS* 値を持つリンクのリストである．select-voted-links では説明に有効な候補と思われる link を選択する．ここでは各事例が max-votes 票を持ち，それを自らが支持するリンク中で *BSS* 値が大きいものから順に投票する．１票でも票を得たリンクのみを, ルール群選択の対象とする. この結果, 扱うリンクの数を少数に限ることとなり，以降の計算が簡単化される．なお投票に当たっては，各事例から同一の主条件を有するリンクへの投票を禁止し，似かよったリンクばかりが選択されることを防いでいる．

投票で選ばれたリンク群（voted-links）から１群のルールを select-rules で選択する．ここで１回の select-rules の実行が，１個のルールセットを生成する．この際，*eBSS* 値が高い順にルールとして選択する．この値は *BSS* 値に，未だ他のルールの支持に利用されていない事例の割合を乗じたものである．これにより，少数のルールで全事例を説明するようなルール群の表現が得られる．

第１番目のルールセットを調べることにより，目的変数識別のための大まかなデータの傾向を知ることができる．また，２番目以降のルールセットには，異なった視点からの識別ルールが載せられており，これらから有効な知識が得られる場合も多い．

しかし，ルールを精細に調査しようとすれば，以下のような問題点が存在する．

(a) 前提条件のみが若干変更されたルールが出現．
(b) 異なったルールセットに属するルール間では支持事例群の重なりが不明．
(c) 多数のルールがその間の関連性が不明なまま記述される．

(3) Datascape を求めて

ルールの構造化

本節では，上記の問題点を解消しdatascape の概観を与える目的で，Algorithm 2 に示す方法によりルール群を選択する．ここでは，各ルールを最適化後，尾根筋情報 (ridges), 関連ルール群情報 (relatives) を付加し，各ルールを構造化して表現する．従って前節のルールセット概念は存在しない．ここで取り上げた３種の改良について以下説明する．

```
create-rule-sets(links)
  voted-links := select-voted-links(links)
  loop changing rset from 1 until max-rset
    rules(rset) := select-rules(voted-links)
    voted-links -= rules(rset)
    print(rules(rset) rset)

select-voted-links(links)
  links := sort-by-BSS(links)
  loop for link in links
    employed := nil
    loop for case in supporting-cases(link)
      if  #votes(case)<max-votes and
          added-item(link)∉votes(case)
        push added-item(link) to votes(case)
        employed := t
    if employed=t
      push link to voted-links
    if #votes(case)<max-votes in all cases
      return voted-links

select-rules(voted-links)
  rules := nil
  loop for link in voted-links
    copy BSS & cases to eBSS & eCases of link
  loop until null(voted-links)
    voted-links := sort-by-eBSS(voted-links)
    rule := pop(voted-links)
    if #rule(cases)=0
      return rules
    loop for link in voted-links
      ecases(link) -= ecases(rule)
      eBSS(link) := (ecases(link)/cases(link))
                    *BSS(link)
    push rule to rules
```

Algorithm 1

ルールの最適化

ルールの主条件と前提条件を構成する節が複数の値を保持できるようにし，これら条件部の値域を変えながら *BSS* が極大値を取るように greedy に最適化する．ここで最適化する対象は，(1) 主条件に現れる変数, (2) 前提条件に現れる変数群, (3) 前記以外の説明変数群, のすべてにわたって行う．この結果，新たに前提条件が付加される場合や，反対に既存の前提条件が削除される場合がある．なお，trivial な前提条件が現れないようにするため，5%以上 *BSS* 値を増大させない前提条件節は最適化後に削除した．

```
create-structured-rules(links)
  ls := sort-byBSS(select-voted-links(links))
  rules := nil
  final-rules := nil
  loop until null(ls)
    rule := pop(ls)
    rule := optimize(rule)
    unless member(rule rules)
      rule.ridges := add-ridges(rule)
      rule.relatives
          := add-relatives(rule rules)
      push rule to rules
  rules := sort-byBSS(rules)
  loop for rule in rules
    push rule to final-rules
    remove rule.relatives from rules
  return final-rules
```

Algorithm 2

強いルールに前提条件が付加されてできた弱いルールは，この操作により同一のルール表現に収束し，結果としてルール数が減少することを期待できる．

ルールの尾根筋

山塊の見取り図が，主たるピークとその周囲の尾根筋によって表されていると，全体の地形を理解しやすい．前節の最適化されたルールは，右辺の変数値識別のためのピーク位置を表す．それでは尾根筋とは何であろうか？ここでは「ルールの支持事例群が変

化しても，*BSS* 値の減少程度が比較的に小さいような前提条件の変化」と定義しよう．

たとえば，ルールの前提条件に新たな節を付加した時，その変数がすでにルール表現中に存在する変数群と特に相関を持たない場合は，カバーする事例数の減少に応じて，*BSS* 値も減少する．しかし，事例数は 40%減少するが，よりシャープなパターンを導くために，*BSS* 値の減少は 10%に留まる場合がある．このような前提条件は，主たるルールをより深く理解させてくれる．反対に，ルールの前提条件の値域を広げることにより事例数が増加する場合でも，*BSS* 値の減少程度が少ない場合がある．

ここでは，支持事例群が一定程度以上（default: 30%）増減した場合，あるいは前提条件節が削除された場合で，かつ *BSS* 値の減少が少ないような条件節を，尾根筋上の点であると考える．このような条件表現を *BSS* 値の減少が少ない順にいくつか提示することにより，ルールの理解に資することとする．

関連ルール群

主条件適用前後の支持事例群がほぼ重なる２種のルールは，実体としては同一の事象の違った側面を表すものである．これらは *BSS* 値の大きい基準ルールと，それに対する関連ルールとして表示することが，利用者にとって便利である．２種のルールがまったく異なった説明変数で表されていても，これらの間の相関が非常に高い場合は，これらを関連ルールとして同時に把握できることになる．

それでは，主条件適用前の支持事例群は大幅に異なるが，適用後は重なる場合をどう考えたら良いであろうか？さらに，主条件適用後の事例群の重なり程度は小さくとも，これらの重なりが近似的にでも部分集合関係にある場合はどうであろうか？

著者らは，これらの場合すべてが関連するルール対であると考え，*BSS* 値が大きい方を基準ルールとし，他方を関連ルールとして表示する．関連ルール選択の規範としては，２種のルール A, B の主条件適用後の支持事例で，(4)式の値が利用者の指定する関連度（default: 0.7）を上回るときに，関連ルールと位置づけた．また，関連ルール出力の順序は主条件適用前の事例群で，(5)式の Tanimoto 係数が大きい順とした．

$$\max\left(\frac{\sup(A\cap B)}{\sup(A)},\frac{\sup(A\cap B)}{\sup(B)}\right) \tag{4}$$

$$\sup(A\cap B)/(\sup(A)+\sup(B)-\sup(A\cap B)) \tag{5}$$

結果および考察

マイニングの例題によく取り上げられる Meningoencephalitis 診断問題（bacteria 性／virus 性の判別）に，前節で述べたルール表現を適用する[8]．この問題でパラメータ群（thres=.05, min-support=.01, thr-BSS-print=.02）を指定して計算すると，48,675 節点のラティスが生成され，その中から 250 個のリンク（BSS>2.8）がルール候補となる．旧来の方法では 25 個のルールが出力される（max-votes=3, max-rule-sets=3）．これらすべてを子細に検討するのは困難であり，また，パラメータ値を変更すると，ルール数も変わるという不安定さも問題となる．

新たに導入したルール出力を, 図２に示した例に沿って解説する. 第１行はこのルールの主条件適用前後の事例数と *BSS* 値を示し，第２行が主条件と前提条件を示す．３行目が目的変数である Diag2 の *BSS* 値と分布の変化を表す．この場合[Cell_Poly: 4]の条件により Bacteria 性の確信度が 29%から 100%に上昇していることが判る．次の２行は，目的変数以外で主条件適用により分布が大きく変わる変数について，同様の情報を記す. ここでは，CSF_CELL の値が高い方へ大きく変化していることが判る.

Ridge information は尾根筋の情報であり，５種の前提条件節を付加した場合の *BSS* 値が示されている．たとえば，AGE が 0 – 3 のカテゴリーにあるという前提条件を付加した場合，主条件適用前後の事例数はそれぞれ 79%, 70%に減少するのに対し，*BSS* 値は 12.0 であるから 78%にしか減少しないことが判る.

Relatives information は関連ルール群情報であり，ここではその最初のもののみを示す．最初の２行は支持事例群の類似性を総括的に示す Tanimoto 係数，および事例群間の重なり数（R&S), 関連ルールのみを支持する事例数（R-S), 反対に基準ルールのみを支持する事例数（S-R）を，主条件適用の前後で示す．後は通常のルール表現が続く.

この例では，主条件適用前で関連ルール 111 事例中の 95 事例（86%）が，また適用後は 16 事例中の 14 事例（88%）が基準ルールと重なっている．この結果から，基準ルールとまったく異なる条件部を持つこの関連ルールは，基準ルールで説明される事例中の半数足らずを，異なる視点から説明するものであるといえる．言い替えれば，この関連ルールは，基準ルールで主条件適用後の 30 事例中のサブクラスター14 事例の特徴を示したものとなっている.

ラティス生成のパラメータ（thres, min-support）を変更して，この問題を解析した場合に，生成されるルール数がどのようになるかを，表１に示す．ここで，括弧内は関連ルール数の合計を示す．なお，他のパラメータは thr-BSS-print=.01, max-votes=3 を採用

```
Rule 1:    Cases: 119 -> 30; BSS=  15.3
  IF [Cell_Poly: 4]   added on  [CSF_CELL: 1 - 4]
  THEN Diag2:       BSS:15.3       0.29 0.71 ==>  1.00 0.00
  THEN CSF_CELL:    BSS:4.96       0.00 0.18 0.18 0.27 0.37 ==> 0.00 0.00 0.00 0.13 0.87
  THEN Cell_Poly:   BSS:10.5       0.19 0.15 0.22 0.18 0.25 ==> 0.00 0.00 0.00 0.00 1.00
Ridge information
 NEWPRECOND:   [CSF_PRO: 0 - 3]      BSS: 12.3    up-cover: 0.857 low-cover: 0.700
 NEWPRECOND:   [AGE: 0 - 3]          BSS: 12.0    up-cover: 0.790 low-cover: 0.700
 NEWPRECOND:   [ESR: 0 - 1]          BSS: 11.8    up-cover: 0.832 low-cover: 0.667
 NEWPRECOND:   [Cell_Mono: 2 - 4]    BSS: 11.7    up-cover: 0.697 low-cover: 0.800
 NEWPRECOND:   [FOCAL: 0]            BSS: 11.5    up-cover: 0.773 low-cover: 0.700

**Relatives information for Rule 1
  [Upper] Tanimoto: 0.70  R&S:   95  R-S:   16 S-R:  24
  [Lower] Tanimoto: 0.44  R&S:   14  R-S:    2 S-R:  16
Rule 1-1:   Cases: 111 -> 16; BSS=  7.716
  IF [CRP: 3]    added on   [FEVER: 0 - 3] [SEIZURE: 0]
  THEN Diag2:       BSS:7.50       0.32 0.68 ==> 1.00 0.00
  THEN STIFF:       BSS:1.47       0.23 0.05 0.36 0.16 0.19 ==> 0.19 0.00 0.19 0.06 0.56
  THEN WBC:         BSS:2.82       0.09 0.38 0.22 0.14 0.18 ==> 0.00 0.00 0.19 0.19 0.62
  THEN CRP:         BSS:8.62       0.54 0.20 0.12 0.14 ==> 0.00 0.00 0.00 1.00
  THEN CSF_CELL:    BSS:1.09       0.14 0.12 0.14 0.23 0.37 ==> 0.12 0.00 0.00 0.19 0.69
  THEN Cell_Poly:   BSS:3.89       0.25 0.15 0.18 0.16 0.25 ==> 0.12 0.00 0.00 0.00 0.87
Ridge information
 PRECOND:      [FEVER: 0 - 4]        BSS: 6.11    up-cover: 1.189 low-cover: 1.437
 NEWPRECOND:   [HEADACHE: 0 - 2]     BSS: 5.87    up-cover: 0.703 low-cover: 0.687
 NEWPRECOND:   [AGE: 0 - 3]          BSS: 5.75    up-cover: 0.820 low-cover: 0.625
```

図２．ルール出力例

し，*BSS* 値が 4.2 以上のルールのみを出力している．探索範囲の広いパラメータ値が多くのルールを与えているが，新しく検知されたものを除けば，すべて同一のルール表現に収束しており，この方法が安定な結果を与えることが判る．

表１．ラティス生成パラメータの影響

min-sup	thres			
	0.03	0.05	0.07	0.10
0.01	9 (12)	8 (11)	6 (7)	4 (6)
0.02	9 (12)	8 (11)	6 (7)	4 (6)

表２．関連性規範の変更による影響

ridge-ratio			
0.9	0.8	0.7	0.5
15 (4)	13 (6)	8 (11)	3 (16)

次に(4)式で指定する関連ルールの規範（ridge-ratio と呼ぶ）を変更した場合に，どのようなルール群の変更が起こるかを調べよう．パラメータ値として thres=0.05, min-sup=0.01 で得られるラティスを例に取ると，多くのルール間で互いに支持事例群の重なり合いが様々な程度に起こっているため，関連ルールへの組織化の程度も，表２のように変わってくる．

次に, max-votes, thr-BSS-print の両パラメータを変化させた場合のルール数を表３と４に示す．少なくともこの例題においては，votes 数はまったく出力結果に影響を与えない．しかし，thr-BSS-print の値は出力ルール数に大きな影響を与える．後者の値は，ラティスから候補リンクを選択する際の *BSS* 値の基準を定めるものである．検出限界を下げて多数の候補リンクを採用すると，結果として非常に多数のルールが生成されることが判る．

表３．Votes 数変更の影響

thr-BSS-print	max-votes			
	1	3	5	10
0.02	8 (11)	8 (11)	8 (11)	8 (11)

表４．候補リンク検知限界変化の影響

max-votes	thr-BSS-print		
	0.01	0.02	0.03
3	15 (108)	8 (11)	2 (3)

考察

少数の構造化されたルール群を得るという目的を達成することができた．しかし，前節の最後に指摘した thr-BSS-print の値を変化させた場合のルール数変化の結果は，新たな２種の問題点を浮かび上がらせている．第１の問題点は，またしても多すぎるルールの問題であり，多数の関連ルールから価値の低いルールを除去するさらに精細な作業が必要となる．この課題の中には支持事例群や条件部のより詳細な比較から解決することのできる部分と，領域固有のオントロジーに依存する部分が存在するであろう．特に，関連ルールの支持事例群から基準ルールと重なる部分を削除した場合，残りの事例が識別能力を持たなければ，その関連ルール自体は無意味なものと考えることができる．この考え方に従って，今後さらにルール数の削減を試みる予定である．

第２の点は，不十分な詳細度でしかラティスを生成できない場合であっても，作成可能なラティス内のリンクから出発して，強い有益なルールを見いだすことが可能な点である．*BSS* を多峰性関数として見た場合，カスケードモデルで検出されたリンクを出発点として単純な山登り法により，多数の極大値へ到達できることを示している．実用面から考えて，重要な貢献をする可能性がある．

参考文献

[1] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules, *Proc. VLDB*, 487-499 (1994).
[2] Okada, T.: Finding Discrimination Rules using the Cascade Model. *J. Jpn. Soc. Artificial Intelligence*, 15, 321-330 (2000).
[3] Okada, T.: Rule Induction in Cascade Model based on Sum of Squares Decomposition, *Proc. PKDD-99*, 468-475, LNAI 1704 (1999).
[4] 岡田孝: "カスケードモデルによる識別ルールの発見と Datascape", 人工知能学会第１３回全国大会, 11-01 東京 (1999).
[5] Gini, C.W.: *Studi Economico-Giuridici della R. Universita de Cagliari* (1912). Reviewed in R.J. Light, B.H. Margolin, *J. Amer. Stat. Assoc.* 66, 534-544 (1971).
[6] Okada, T.: *Kwansei Gakuin Studies in Computer Science*, 14, 1-6 (1999). (URL = http://www.media.kwansei.ac.jp/home/kiyou/kiyou99/ kiyou99-e.html)
[7] Okada, T.: Efficient Detection of Local Interactions in Cascade Model, Proc. PAKDD 2000, 193-203, LNAI 1805 (2000).
[8] Washio, T.: http://wwwada.ar.sanken.osaka-u.ac.jp/pub/ washio/jkdd/jkddcfp.html

２．変異原性および発ガン性に関連する化合物特徴の認識

検討内容

カスケードモデルによる解析が，実際に化合物の生理活性を認識するためにどの程度役立つかを調査することが必要である．このために以下の２種のデータを対象として実際に解析を遂行した．

(1) 芳香族ニトロ化合物の変異原性データ
(2) ネズミに対する化学発ガン性の調査データ

この際，説明変数としての化学構造は，構造式のグラフから生成される多数のフラグメント群が存在するか否かで表現すると共に，LogP 値（水・オクタノール分配係数）や HOMO 値などの代表的な物理化学的性質をも取り入れた．

結果及び考察

変異原性に対する適用結果から，「芳香族ニトロ化合物で ortho 位置換基の有無が重要な因子となる」というような有意義な結果が得られた．また，発ガン性への適用結果からは，「有機塩素化合物の活性と水素結合受容体の有無および分子の柔軟性との間の高い相関」など，数多くの有用な知見を得ることができた．これらの解析結果は，化学構造の詳細にわたる議論を含むため，本報告の末尾に付録として下記２報の論文を付す．

(1) T. Okada: Discovery of structure activity relationships using the cascade model: the mutagenicity of aromatic nitro compounds, *J. Computer Aided Chemistry*, Vol.2, 79-86 (2001).

(2) T. Okada: Characteristic substructures and properties in the chemical carcinogenicity studied by the cascade model, Proceedings of the international workshop on predictive toxicology challenge 2001, Freiburg (2001).

発ガン性の解析と予測は，国際ワークショップ Predictive Toxicology Challenge 2001 に参加して発表したものである．このワークショップでは，主催者により参加者のモデル評価が行われた．以下にワークショップの概要とわれわれの結果への評価を紹介する．

ワークショップは次ぎに示す３段階を経て行われた．

(1) Data engineering: NTP により調べられた化合物417種に対する発ガン性データを主催者が整理して公開し，これらに対する記述子を一般から募集した．その結果 7000 以上におよぶ記述子群が利用可能となった．筆者はそれらの中から Kramer による線形 fragment と TReimers による物理化学的性質の一部を利用した．

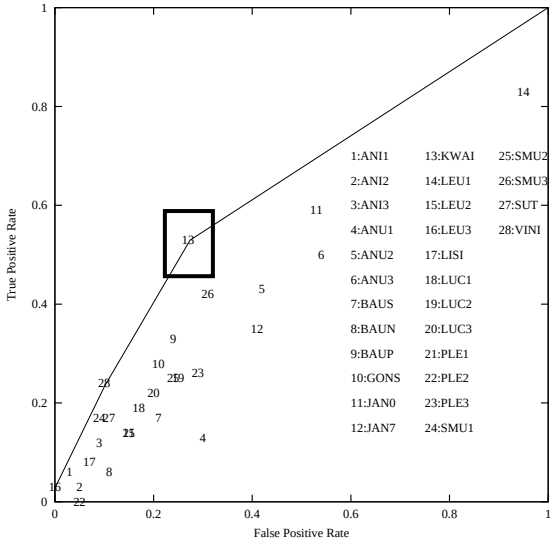


図1　雌のratに対するROC結果

(2) Model construction: 上記の記述子群を利用して，参加者が自らのモデルを作成し，試験結果を伏せた化合物群 185 種（FDA による試験結果）に対して，その発ガン性を予測する．この段階では 14 の研究グループが 26-31 のモデルを提出した．これらモデルの基盤となった方法論は，決定木，帰納論理プログラミング，ラティスの被覆探索，support vector machine，ニューラルネット，通常の回帰，およびこれらの複合的方法など多岐にわたる．

(3) Model evaluation: テスト用化合物群について各方法で得られた結果を集約し，主催者が ROC 分析によって評価した．Female rat に対する結果を，図１に示す．四角で囲んだ点が筆者のモデルである．ROC では左上に存在するモデルほど良い予測であるから，筆者の結果は非常によい評価を得ることができた．さらに，US NIEHS と EPA の専門家が理解のしやすさとその毒性研究における有用性の観点から，各レポートを評価した．筆者の報告は総合的にみて最良のものであるとの評価を与えられた．

まとめ

現段階のカスケードモデルによる解析により，変異原性と発ガン性に関して実用レベルの構造と活性の相関に関する知識を得ることができた．ここでの問題点は，ルール群の解析に労力がかかる点にあるといえる．他方，第１節に述べたように，ルールを最適化し，さらに尾根筋情報と関連ルール情報を付加して，ルールを構造化すれば，安定でかつ冗長度の低いルール群表現を得ることができる．

今後この新しいルール群表現を用いて，各種の化学構造と生理活性データを解析し，本研究の目的を達成する計画である．なお，対象データとしては１２万種におよぶ治験薬品のデータベース，および２１万種におよぶ毒性データベースをすでに整備した．今後，多数の生理活性について解析を行い，その結果を WWW によって公開していく予定である．

Discovery of Structure Activity Relationships using the Cascade Model: The Mutagenicity of Aromatic Nitro Compounds

Takashi Okada *

Center for Information & Media Studies, Kwansei Gakuin University
1-1-155 Uegahara, Nishinomiya, Hyogo, 662-8501 Japan

(Received September 29, 2001; Accepted October 29, 2001)

The cascade model is a rule induction methodology that uses the level-wise expansion of a lattice. An attribute-value pair is expressed as an item, and every node in the lattice is specified by an itemset and its supporting instances. If the distribution of the class attribute values suddenly changes along a link in the lattice, the link is represented as a rule "IF *item-along-link* added on *itemset-on-upper-node*, THEN *class-i*". The strength of the rule is measured by the *BSS* value of the link. In the SAR (structure activity relationship) study, we generate many linear substructure patterns like "NH-C-C-C-OH" from the data set of molecular structures. Matching between a pattern and a molecule leads to an item [pattern: y or n] depending on whether the pattern exists in the molecule. Application of the cascade model to this item data set gives us SAR rules. The resulting rules are examined by referring to the supporting molecular structures. Several rules have led to valuable working hypotheses, including the importance of steric hindrance to the coplanarity of NO_2 to the activity level.

Key Words: data mining, cascade model, fingerprint, SAR, mutagenicity, aromatic nitro compounds

1. Introduction

The SAR (structure activity relationship) between chemical structures and biological activity has always been an important research field in chemistry and biology. Recent innovations in high throughput screening technology have resulted in a vast amount of SAR data, and new data mining technology is expected to facilitate the drug development process. The principal aim of this paper was to discover SAR rules from the mutagenicity data of some chemical compounds, but we also expect that the mining process used here will be widely applicable to the qualitative SAR recognition problem.

The KDD Challenge 2000 Workshop was held to bring together researchers and practitioners interested in

* *okada@kwansei.ac.jp*

discovering knowledge from real-world databases.[1] One of the target datasets was the mutagenesis activity of 230 aromatic and heteroaromatic nitro compounds compiled by Debnath et al.[2] Their structures and various descriptors were provided by the author (also included in the supplementary materials of this paper).

We analyzed the data set to obtain rules that help in understanding the mutagenicity of chemical compounds at the workshop.[3] The descriptors used were limited to those directly derived from structural formulae. In this paper, we add some physicochemical properties to the descriptor set, and inspect the resulting rules referring structures of the supporting compounds.

The basic mining scheme consists of the following three steps. First, we generate thousands of items from a set of molecular graphs, where each item denotes whether a specific linear substructure pattern is contained in a molecule. Then, we describe each molecule as a set of items like [pattern-1: y, pattern-2: n, ...]. Some items representing activity and physicochemical properties are also added. The second step is to apply the cascade model to mine SAR rules. The method can detect specific combinations of substructure patterns leading to high or low mutagenicity. The last step is to polish the rule expression by consulting the compound database so that biologists and chemists can evaluate the results.

Section 2 briefly introduces the cascade model used for the mining process; Section 3 explains the item generation scheme used. The computation procedure and the results obtained for the challenge problem are shown in Section 4, along with the scheme of rule interpretation. Section 5 compares the current results with previous works, and discusses possible improvements in the mining process.

2. The Cascade Model

The model examines an itemset lattice in which an [attribute: value] pair is used as an item to form itemsets. Links in the lattice are selected and expressed as rules.[4] Figure 1 shows a typical example of a link and its expressed rule. Here, the problem contains five attributes: A - E, each of which takes (y, n) values. The itemset at the upper end of the link has item [A: y], and another item [B: y] is added along the link. Items of the other attributes are called veiled items. The tables attached to the nodes show the frequencies of veiled items.

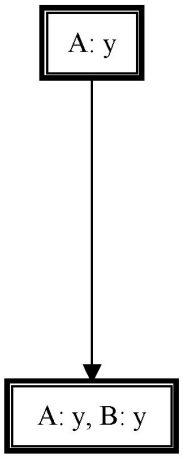
To evaluate the strength of a rule, the within-group sum of squares (*WSS*) and between-group sum of squares (*BSS*) are defined by the following formulae,[5,6]

$$WSS_i = \frac{n}{2}\left(1 - \sum_a p_i(a)^2\right) \quad , \tag{1}$$

$$BSS_i = \frac{n^{\mathrm{L}}}{2}\sum_a \left(p_i^{\mathrm{L}}(a) - p_i^{\mathrm{U}}(a)\right)^2 \quad , \tag{2}$$

where i designates an attribute, the superscripts U and L indicate the upper and lower nodes, respectively, n is the number of cases supporting a node; and $p_i(a)$ is the probability of obtaining the value a for attribute i.

Figure 1 shows the WSS_i and BSS_i values along with their sample variances. A large BSS_i value is evidence of a strong interaction between the added item and attribute i. The textbox on the right in Fig. 1 shows the derived rule. The added item [B: y] appears as the main condition on the LHS, while the items in the upper node are placed at



	y	n	*WSS*	σ^2
B	60 (9.6)	40 (14.4)	24.0	.24
C	50 (12.5)	50 (12.5)	25.0	.25
D	60 (9.6)	40 (14.4)	24.0	.24
E	40 (14.4)	60 (9.6)	24.0	.24

	BSS
B	9.60
C	0.00
D	6.67
E	5.40

IF [B: y] added on [A: y]
THEN [D: y; E: n]
Cases: 100 → 60
[D: y] 60% → 93%, BSS = 6.67
[E: n] 60% → 90%, BSS = 5.40

	y	n	*WSS*	σ^2
B	60 (0.00)	0 (0.00)	0.00	.000
C	30 (7.50)	30 (7.50)	15.00	.250
D	56 (0.25)	4 (3.48)	3.73	.062
E	6 (4.86)	54 (0.54)	5.40	.090

Fig. 1 A sample link, its rule expression and properties of the veiled items.

the end of the LHS as preconditions. When a veiled attribute has a large BSS_i value, one of its items is placed on the RHS of a rule.

We can control the appearance of attributes on the LHS by restricting the attributes in the itemset node. On the other hand, the attributes on the RHS can be selected by setting the minimum BSS_i value of a rule ($min\text{-}BSS_i$) for each attribute. Items on the RHS of a rule do not need to reside in the lattice. This is in sharp contrast to association rule miners, which require the itemset, [A: y; B: y; D: y; E: n] to derive the rule in Fig. 1. These characteristics of the cascade model make it possible to detect rules efficiently.[7]

However, the most important problem when we use a mining method based on lattice expansion is the combinatorial explosion in the number of nodes,. The pruning method based on the constraint of *BSS* values can overcome this problem to a certain extent.[8] This is especially important in a problem with thousands of attributes. The cascade model was implemented as DISCAS software using lisp, and is used in this work.

3. Item Generation from Graphs

Our objective is to provide an itemset expression of a molecular structure. The items generated are handed to DISCAS, which gives us SAR rules. The items do not need to restore the original structural formula. However, a chemist needs to understand the meanings of the items, and must be able to use the resulting rules to guide the process of molecular design. Since some specific substructure is often necessary for a biological activity to appear, it is natural for an item to have a meaning, whether or not a substructure pattern exists in a molecule. That is, we must provide understandable fingerprints.

The number of all possible substructure patterns in a set of molecular graphs is usually too large to be used in the analysis. We introduced the method of relative indexing of vertices to limit the number of patterns.

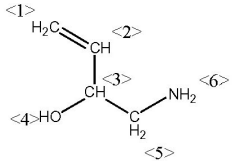
3.1. Relative indexing of vertices

This method was originally proposed to generate items from syntactic parse trees,[9] but has been extended to provide linear connected subgraph patterns from chemical structural formula. Figure 2 shows an example of a structural formula and a set of substructure patterns derived from it. Hydrogen atoms are regarded as attachments to heavy atoms. Here, every pattern consists of two terminal atom parts and a connecting part along the shortest path. For example, the pattern at the bottom of the left column uses atoms <1> and <6> as terminals, and the connecting part is described by "=C-C-C-", which shows the sequence of bond types and element symbols along the path, <1>=<2>-<3>-<5>-<6>. An aromatic bond is denoted as "r". The description of a terminal atom includes the coordination number (number of adjacent atoms), as well as whether there are attached hydrogen atoms.

In this example, we require that at least one of the terminal atoms is a heteroatom or an unsaturated carbon atom. Therefore, no pattern appears between tetrahedral carbon atoms <3> and <5> in Fig. 2. Patterns consisting of a single heavy atom like C3H and O2H have also been added to the items. The itemset based on these patterns can be regarded as constituting of a kind of fingerprint of the molecule, and is similar to the descriptor set Klopman used in the CASE system.[9]

3.2. Substructure pattern generation using various schemes

The substructure patterns in Fig. 2 are just examples. The coordination number and/or hydrogen attachment in the terminal atom part may be omitted. The connecting part is also subject to change. For example, we can cut element symbols or bond types, and use just the number of intervening bonds as the connecting part. Conversely, we can add coordination numbers and/or the hydrogen attachment to the atom descriptions in the connecting part. There is no *a priori* criterion to judge the quality of various substructure pattern expressions. Only after the discovery process can we judge which type of pattern expression is useful.



C3H=C3H	C3H-C4H	C4H-N3H
C3H=C-C4H	C3H-C-O2H	C4H-C-N3H
C3H=C-C-O2H	C3H-C-C4H	N3H-C-C-O2H
C3H=C-C-C4H	C3H-C-C-N3H	C4H-O2H
C3H=C-C-C-N3H		C4H-C-O2H

Fig. 2 A structural formula and the derived linear substructure patterns.

Table 1 Effects of Item Generation Schemes

Terminal atom part	Connecting part								
	Element names & bond types			Bond types only			Number of bonds and terminal bond types		
Coordination no. & hydrogen attachment	2044 → 77;	10.7	(7)	1710 → 85;	13.0	(8)	822 → 72;	14.1	(10)
Coordination no.	1678 → 46;	9.8	(7)	1198 → 59;	14.2	(9)	531 → 47;	11.4	(4)
Element name only	1587 → 51;	8.5	(7)	1062 → 57;	14.2	(10)	412 → 51;	11.6	(5)

Each cell depicts the *number_of_generated_items* → *number_of_items_used*; *sum_of_squares_explained* (*number_of_rules_in_the_first_rule_set*).

As a preliminary survey to check the usefulness of various expressions, we used 3 schemes for the terminal atom part, and combined them with 3 schemes for the connecting part. The names of the columns and rows in Table 1 show this scheme used for item generation.

We applied this item generation scheme to the mutagenicity data set described below. The number of generated patterns ranged from several hundred to more than two thousand. Most of the patterns appear in a few cases or in almost all cases. Such patterns with an unbalanced distribution are automatically omitted in the actual computation process of DISCAS, because the pruning condition always prevents such items from entering the itemset of a lattice node.

Each cell in Table 1 depicts the result of computation from a combination of the row and column schemes. The first number to the left of "→" shows the number of substructure patterns generated. The second number to the right of "→" shows the number of patterns actually used in the cascade model calculations. Preliminary cascade model calculations result in a set of rules that does not use items derived from LogP or LUMO. The number of rules in the first rule set and the sum of squares value explained by these rules are shown to the right of each cell. Here, we set the pruning parameters to *minsup* = 0.1 and *thres* = 0.1.

We can see that the number of patterns generated increases as the pattern description goes into detail. However, there is no general tendency in the number of patterns used or in the explained sum of squares values. When we inspect the resulting rules generated by the cascade model, rules using a simpler item description are very hard to interpret, as the item expressions are very different from those in the language of chemistry. On the other hand, we can easily associate a rule with the structures of the supporting compounds, if the item is expressed in detail. Therefore, we use the item generation scheme shown in the upper left corner of Table 1 in the rest of this paper. This leads to linear substructure patterns, like those in Fig. 2.

4. Results and Discussion

4.1. Details of computation

Debnath et al. compiled mutagenicity data on 230 aromatic and heteroaromatic nitro compounds.[2] The SDF dataset in the supplementary materials is used for the SAR study (see attached files). Each record contains a chemical graph with activity, LogP, and LUMO values attached.

The cascade model cannot treat continuous variables directly. Therefore, we categorized numerical attributes by visually inspecting histograms using splitting values: activity (-90, 0, 3); LogP (2, 4, 6); and LUMO (-2.5, -1.5). The four categories of activity are inactive, low, medium, and high, and their respective percentages are 15.2, 44.3, 30.9, and 9.6%. The *BSS* value of the categorized activity was calculated assuming that each category is nominal.

The item generation process provided 2,044 substructure patterns, as depicted in Table 1, and the items corresponding to these patterns constitute a feature vector for every molecule. The item dataset was analyzed using DISCAS software, with the pruning conditions set to *minsup* = 0.05 and *thres* = 0.1. Seventy-seven substructure patterns remained to be used for this pruning condition. DISCAS generated a lattice containing 1, 98, 2,240, 6,324 nodes at lattice levels with 0–3 items, respectively. It took 147 seconds to create all 8,663 nodes using a PC equipped with a 266-MHz Pentium II.

A link was selected as a candidate rule if its *BSS* exceeded 2.3 (1% of the number of instances). There were 229 candidate links, and they were expressed using three rule sets to facilitate inspection by the user. The rules in a rule set were selected so that their supporting instances did not overlap and the rules were maximally independent of each other. The first rule set contained 11 rules, and these are expected to be the most interesting; the second and third rule sets contained 14 and 13 rules, respectively.

Two rules can be regarded as different sides of the

same reality if they share the same supporting instances. Although there appeared to be many rules in the second and third rule sets, their supporting instances were the same as for some of the rules in the first rule set. Therefore, the results shown in the following subsections are drawn from the rules in the first rule set.

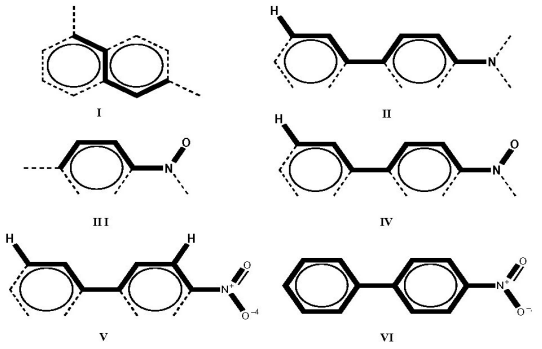
The *BSS* value of the strongest rule is larger than that in the previous work,[3] and we can judge that the current results can give us deeper insight for SAR recognition.

4.2. Interpretation of rules

Rules derived from itemset representation are not directly understandable. For example, the third rule in the first rule set is

IF [C3HrCrC-CrCrCrC-N3: y]
added on [C3rCrCrCrC3: n]
THEN Activity = low
40.8% → 14.0%; BSS: 3.25; Cases: 157 → 43
0.10 0.41 0.41 0.08 ⟶ 0.00 0.14 0.58 0.28

The precondition states that there are never 4 consecutive aromatic bonds as in **I**, while the main condition reveals the importance of substructure **II**. The RHS indicates that there is a large decrease in the percentage of compounds with [Activity=low]. The fourth line of this rule shows that only 43 of the 157 compounds selected by the precondition satisfy the main condition. The percentage of [Activity=low] decreases from 40.8 to 14.0%, and the *BSS* value of this rule is 3.25. The bottom line shows the detailed distribution of the activity levels (*inactive*, *low*, *medium*, *high*) for the



compounds before and after applying the main condition. In this rule, we can see that the main condition has shifted the distribution to higher activity levels.

DISCAS can write optional RHSs on request. A sample of an optional RHS for the above rule is shown below. An attribute-value pair and its change in percentage are depicted if it has high correlation with the main condition.

THEN C3rCrCrC-N-O1 = y
68.2% → 100.0%; BSS: 4.36

This substructure pattern is shown as **III**. As its percentage becomes 100% after applying the main condition, the conjunction of patterns **II** and **III** should be regarded as the real main condition. Retrieval of the data set showed that these patterns can be unified to give a larger pattern, **IV**. Consideration of other optional RHSs led us to conclude that **V** should be the substructure pattern for the main condition of this rule.

The retrieval of substructure **V** showed that all 43 compounds supporting this rule share substructure **VI**, while none of the excluded 114 compounds contain it. Therefore, this rule can be stated as "After we exclude fused polynuclear aromatic compounds, IF a 4-nitro-biphenyl (**VI**) substructure exists in a compound, THEN it is expected to be more mutagenic than otherwise."

Another interesting interpretation scheme was necessary for the main condition of the following rule.

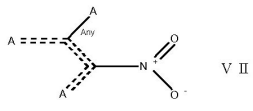
IF [C3HrCrCrC-N3: n]
added on [C3rCrCrC-N-O1: n]
THEN Activity = medium
39.3% → 88.2%; BSS: 3.13; Cases: 56 → 17
0.18 0.43 0.39 0.00 → 0.00 0.12 0.88 0.00

Nitrogen atoms in the pre- and main conditions mean an NO_2 group. Then, this rule tells us that there is neither a CH group nor a substituted carbon atom at the *para* position from the NO_2 group in an aromatic ring. Actually, NO_2 substitution occurs at the nonaromatic 5-membered rings in all 17 compounds supporting this rule. Therefore, reference to the database is essential to interpret a rule. We used Spotfire structure visualization software to access structural formulae indicated by a rule.

4.3. Structure activity relationships

Table 2 shows the main conditions, as well as the preconditions of the 11 rules in the first rule set. The number of compounds and activity distributions are shown before and after applying the main condition. They are ordered by the effective *BSS* values, defined by *BSS*• *#new_cases*/*#supporting_cases*. Here, a *new_case* means a compound that has not appeared in the supporting cases of the previous rules.

The main condition of the first rule, a high LUMO level, applies to the compounds with substructure **VII** resulting in decreased activity. The same main condition

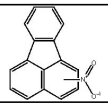


V II

appears in rule 6 without any precondition. That is, the effect of a high LUMO in decreasing the activity is stronger with this precondition.

Figure 3 shows the distribution of the activity concerned with the first rule. Here, the upper and lower

Table 2 Rules Obtained from Aromatic and Heteroaromatic Nitro Compounds

No	Main condition	Preconditions	Characteristics	Percentages of (*inactive, low, medium, high*) / #compounds*	*BSS*
1	[LUMO: high (>-1.5)]	[C3rC-N-O1: y]	See text	(.11 .32 .40 .17) / 128 ⇒ (.20 .63 .17 .00) / 41	3.84
2	[C3HrCrCrC-N3: n]	[C3rCrCrC-N-O1: n]	Nonaromatic 5 membered rings with NO_2 group	(.18 .43 .39 .00) / 56 ⇒ (.00 .12 .88 .00) / 17	3.13
3	[C3HrCrC-CrCrCrC-N3: y]	[C3rCrCrCrC3: n]	*p*-nitrobiphenyl without fused aromatic rings	(.10 .41 .41 .08) / 157 ⇒ (.00 .14 .58 .28) / 43	3.25
4	[C3rCrCrCrC-N3: y]	none	Fused aromatic rings with NO_2	(.10 .31 .44 .15) / 230 ⇒ (.09 .09 .48 .34) / 65	2.69
5	[LogP: low (<2.0)]	none	Substituted nitrobenzenes and heteroaromatic nitro compounds	(.10 .31 .44 .15) / 230 ⇒ (.12 .61 .27 .00) / 51	3.61
6	[LUMO: high (>-1.5)]	none	Rule 1 without its precondition	(.10 .31 .44 .15) / 230 ⇒ (.15 .51 .31 .03) / 99	3.64
7	[C3rCrCrCrC-N3: n]	[C3HrCrCrC-N3: y, C3rCrCrCrCrC3H: y]	No characteristic substructure	(.12 .22 .42 .25) / 69 ⇒ (.14 .71 .14 .00) / 14	2.70
8	[C3HrCrCrCrC-N3: y]	[C3-CrC3H: y]		(.00 .20 .55 .25) / 69 ⇒ (.00 .00 .29 .71) / 14	2.31
9	[LUMO: medium (-2.5:-1.5)]	[C3rCrCrC-N-O1: n]	No characteristic substructure	(.18 .43 .39 .00) / 56 ⇒ (.05 .16 .79 .00) / 19	2.34
10	[C3-CrCrCrC-N3: y]	[C3rCrCrCrC3: n]	*p*-nitrobiphenyl & *p*-nitrophenyl ketone without fused aromatics	(.10 .41 .41 .08) / 157 ⇒ (.00 .19 .55 .26) / 47	2.54
11	[C3rCrCrC-N3: n]	[C3rC-N-O1: y]	*ortho* substituted nitrobenzene with no substituents at the *para* position	(.11 .32 .40 .17) / 128 ⇒ (.22 .59 .19 .00) / 32	2.57

* The left and right sides of the arrow show percentages and #compounds before and after applying the main condition.

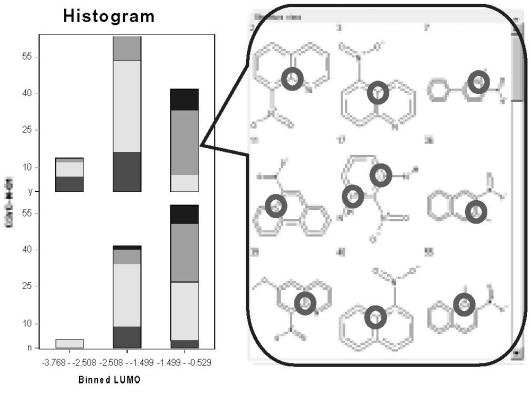


Fig. 3 Activity distributions changing LUMO (x-axis) [Left] and sample compounds with a high LUMO level and substructure **VII** [Right].
Upper chart: compounds with **VII**,
Lower chart: compounds without **VII**,
■: inactive, ▒: low, □: medium, ▓: high.

bar charts correspond to the compounds with and without **VII**, respectively. A bar chart illustrates the distribution of activity levels using LUMO categories as the x-axis. We can see that the effect of a high LUMO level is stronger in the upper chart.

Some of the compounds supporting this rule are depicted to the right in Fig. 3. Red circles indicate substituents at the *ortho* position from NO_2. As they share no characteristic electronic property, we can suggest steric hindrance as the reason for the lower activity. That is, if a substituent violates the coplanarity between an aromatic ring and NO_2, LUMO with a high energy level might lose mutagenic activity. The compound in the upper left corner of Fig. 3 does not have any steric hindrance, and the above hypothesis seems not to apply. However, another molecule may coordinate the nitrogen atom in the pyridine ring and it may cause some steric hindrance.

Table 2 shows the characteristics obtained using other rules. Of these, rules 2, 3, 8, and 11 give us meaningful substructures and narrow activity ranges. Rule 10 shares most of its conditions with rule 3, and the difference in their supporting compounds shows the lower mutagenicity of *p*-nitrophenylketones. These results

should be regarded as material for further studies to give fruitful ideas for toxicologists.

5. Concluding Remarks

Let us compare our results with those obtained by regression and ILP (inductive logic programming) methods. Debnath et al. performed a regression analysis of this data set, and pointed out the importance of LogP and LUMO.[2] They introduced two indicator variables representing characteristic substructures in order to attain reasonable adaptation of regression formula. One was whether three or more fused rings exist, and the other indicates the presence of aceanthrylene substructure. The importance of these fused rings has been recognized in rules 3, 4, 8, and 10 in Table 2. The importance of LogP and LUMO was also detected by rules 1, 5, 6, and 9 in our study. Regression analysis of a set of compounds selected by rules may lead to more precise results using LogP and LUMO.

King et al. used the ILP method to detect substructures affecting the mutagenicity.[11] They included partial charges of atoms calculated by a semi-empirical MO method. Their results contained 6 substructures, four of which used the partial charge to express the characteristic substructures. The resulting substructures contained naphthalene rings, biphenyls, and five-membered rings. Our results include these substructures, and enabled more detailed analysis. Consequently, the current method gives more suitable results for interpretation by chemists and biologists.

The structural diversity of the compounds in the data set forced the division of the compound library in the regression analysis, and this was inherited by the ILP study. However, it is not always easy to divide compounds into several libraries in a reasonable way. The current method treats all compounds uniformly, which is another benefit. Apriori-based graph mining is another method used to analyze this entire data set. Its results also suggested the importance of steric hindrance in nitro benzenes to the mutagenicity.[12]

The scheme employed in this paper was first proposed in our previous paper.[3] Recently, a quite similar approach was employed to analyze HIV data, using linear fragment features and modified association rule mining scheme.[13] Further studies are necessary to judge the adequateness of descriptors and mining schemes employed in these works.

Last, we must note that the pruning condition used in our study is too tight to mine all meaningful rules. Since the DISCAS system simulates a database in memory, the number of nodes in the lattice is limited. As a result, we could not obtain links with fewer supporting instances. For example, the current computation could not perceive 17 compounds with an NH_2 group, although all of them have *low* or *medium* activity levels. Further developments of the DISCAS system are necessary to scale up the computation size.

Part of this research is supported by Grant-in-Aid for Scientific Research on Priority Areas (B) 13131210 and Grant-in-Aid for Scientific Research (B) 12480088.

6. References and Notes

[1] E. Suzuki, KDD Challenge 2000: (URL= http://www.slab.dnj.ynu.ac.jp/challenge2000/).
[2] A.K. Debnath et al., *J. Med. Chem.* **34**, 786–797 (1991).
[3] T. Okada, *Proc. Int. Workshop KDD Challenge on Real-world Data*, 47-53, PAKDD-2000 (2000).
[4] T. Okada, *J. Jpn. Soc. Artificial Intelligence*, **15**, 321-330 (2000).
[5] C.W. Gini, *Studi Economico-Giuridici della R. Universita de Cagliari* (1912). Reviewed in R.J. Light, B.H. Margolin, *J. Amer. Stat. Assoc.* **66,** 534-544 (1971).
[6] T. Okada, *Kwansei Gakuin Studies in Computer Science*, **14**, 1-6 (1999). (URL = http://www.media.kwansei.ac.jp/home/kiyou/kiyou99/ kiyou99-e.html).
[7] T. Okada, *Principles of Data Mining and Knowledge Discovery* (*PKDD'99*), 468-475, LNAI 1704, Springer-Verlag (1999).
[8] T. Okada, *Knowledge Discovery and Data Mining PAKDD-2000*, 193-203, LNAI 1805, Springer-Verlag (2000).
[9] T. Okada, M. Oyama, *Principles of Data Mining and Knowledge Discovery* (*PKDD 2000*), 550-557, LNAI 1910, Springer-Verlag, (2000).
[10] G. Klopman, *J. Amer. Chem. Soc.* **106**, 7315-7321 (1984).
[11] R.D. King, S.H. Muggleton, A. Srinivasan, M.J. & Sternberg, *Proc. Natl. Acad. Sci. USA*, **93**, 438-442 (1996).
[12] A. Inokuchi, T. Washio, T. Okada, & H. Motoda, submitted to *J. Computer Aided Chemistry*, **2**, 87-92 (2001).
[13] S. Kramer, L.D. Raedt & C. Helma, *KDD 2001 Proceedings*, 136-143, ACM (2001).

Supplementary Materials

main.pdf is an introduction to the mutagenicity data set used in KDD challenge 2000. It contains explanations to the following data sets: *okd.sdf*, *okd.csv*, *okd.pdf*, *c2.csv*, *moe.csv*.

Characteristic Substructures and Properties in the Chemical Carcinogenicity Studied by the Cascade Model

Takashi Okada

Center for Information & Media Studies, Kwansei Gakuin University
1-1-155 Uegahara, Nishinomiya, Japan
okada@kwansei.ac.jp

Abstract: The cascade model is a rule induction methodology using the levelwise expansion of the lattice. An attribute-value pair is expressed as an item, and every node in the lattice is specified by an itemset and by its supporting instances. If the distribution of the class attribute values shows a large change along a link in the lattice, the link is represented as a rule "IF *added-item-along-link* added on *itemset-on-upper-node*, THEN *class-i*". The strength of the rule is measured by the *BSS* value of the link. In this study, we utilize linear substructure fragments and several physicochemical properties to describe a rule. A fragment leads to one of the two items [frag-i: y] and [frag-i: n] depending on whether or not the fragment exists in a molecule. Application of the cascade model to these items data set gives us rules about carcinogenicity. We could find several rules with large *BSS* values. Substructures and properties that appear in these rules are expected to provide a starting point for further chemical and biological study. Several rules with classification capability are used to predict the carcinogenicity for the compounds in the test set.

1 Introduction

The importance of SAR (structure activity relationship) study between chemical structures and biological activity is well established in the medicinal sciences. Moreover, the innovation in the high throughput screening technology resulted in the vast amount of SAR data, and a new data mining technology is expected to facilitate the drug development process. The principal aim of this paper is to acquire SAR rules from the Predictive Toxicology Challenge dataset [1] that lead to valuable hypothesis generation for further chemical and biological studies. The process employed here will be widely applicable to the qualitative SAR recognition problem.

We use the cascade model to extract valuable rules. The condition of a rule is represented by the combination of items. An item denotes the existence or the absence of a molecular fragment, or it may be a category of a numerical property.

Section 2 gives a brief introduction of the cascade model. The computation procedure for the challenge problem and its results are shown in Section 3. However, accurate classifications are not the aim of this paper. Rules with large *BSS* values do not always lead to accurate classifications, but they provide interesting viewpoints to analyze the data and to proceed to further research. Some of these rules are referred in Section 3.4. The last section gives a discussion on the preferable improvements in the mining process.

2 The Cascade Model

The model was originally proposed by the author [2]. It can be considered as an extension of the association rule mining. The method creates an itemset lattice where an [attribute: value] pair is employed as an item to constitute itemsets. Links in the lattice are selected and expressed as rules. That is, we watch the distribution of the RHS attribute values along all links, and if there appears a sudden change of distribution along some link, then we bring the two terminal nodes of the link into focus. Think that the itemset at the upper end of a link is [A: y], and an item [B: n] is added along the link. If a sharp activity decrease is found along this link, we can write a rule with the following expression,

IF [B: n] added on [A: y] THEN [Activity: low].

where the added item [B: n] is the main condition of the rule, and the items on the upper end of the link ([A are considered as preconditions. We can put any number of items in the RHS of a rule, if its distribution sho strong interaction with the main condition.

In order to evaluate the strength of a rule, the within-group sum of squares (*WSS*) and between-group su squares (*BSS*) are defined by the following formulae [3, 4],

$$WSS_i = \frac{n}{2}\left(1 - \sum_a p_i(a)^2\right) \quad , \tag{1}$$

$$BSS_i = \frac{n^{\mathrm{L}}}{2}\sum_a \left(p_i^{\mathrm{L}}(a) - p_i^{\mathrm{U}}(a)\right)^2 \quad , \tag{2}$$

where i designates an attribute; the superscripts U and L indicate the upper and lower nodes, respectively; n s the number of supporting cases of a node; and $p_i(a)$ is the probability of obtaining the value a for attribute i. *BSS* takes a large value when the cases at the lower node show exceptional distribution compared to that o upper node. Then, we can set our focus to the main condition part.

The formulation of the model was extended to cover the mining of classification rules and characteristic in a unified framework [5]. When we employ a mining method using the lattice expansion, there always ap the problem of combinatorial explosion in the number of nodes. A new pruning criterion opened a way to with this difficulty [6]. The cascade model was implemented as DISCAS, and it has already been applied t analysis of chemical mutagenicity problem successfully [7].

3 Results and Discussion

3.1 Computation by DISCAS

We need to provide an itemset expression of a molecular structure. The itemset does not need to restore structural formula. However, an expert needs to understand the meaning of items. We employed all li fragments in four class_blind_0.10_fragment_table's by Kramer [1]. The number of fragments is categorized (presence) and 0 (absence). Also included are 9 physicochemical properties given by Treymer [1]. Their name categorization thresholds are CLOGP (0 4), FLEX (0.05 0.25 0.50), VOLUME (150 300), SURF_AREA (150 HBD (0 2), HBA (0 2), LUMO (-0.15 -0.10 -0.05), HOMO (-0.25 -0.20) and Dipole (2 4). We select them, as are easy to understand. Categorizations are simply done by the visual inspection of histograms.

Four datasets for male (female) rat (mouse) were analyzed by DISCAS software, where the pruning condi were set to *minsup* = 0.01 and *thres* = 0.1; their meanings are in [6]. DISCAS generated a lattice containing 4 to 60000 nodes after 7 to 12 minutes using a PC with 450MHz Pentium III. A link was selected as a rule cand if its *BSS* is larger than 1.7 (0.5% of cases). The rule selection process chose 134 - 919 candidate links, and were represented as three rule sets with 10 to 30 rules.

3.2 A sample rule

The strongest rule, the first rule in the first rule set, has the following expression in the application to the mouse data set,

```
IF [HBA = 0] added on [FLEX > 0.5] THEN [MM = p]
     43.0% -> 94.7%; BSS: 5.08; Cases: 79 -> 19   .
```

The precondition means that a molecule is very flexible, while the main condition reveals the absence of hydr bond acceptors. The RHS denotes that a large change is observed in the percentage of compounds with [MM (positive carcinogenicity for male mouse). The second line of this rule denotes that only 19 compounds satisf main condition among 79 compounds selected by the precondition. The percentage of [MM = p] increases 43.0% to 94.7%, and the *BSS* value of this rule is 5.08.

Figure 1 shows a pie chart illustrated by Spotfire software. It shows the distribution of positive and neg cases for categorized HBA (y-axis) and FLEX (x-axis) values. Red and blue areas denote positive and neg cases, respectively. Y-axis categories are HBA=0, 1=HBA=2, and HBA=3 from the bottom. X-axis categorie FLEX=0.05, 0.05<FLEX=0.25, 0.25<FLEX=0.5, and FLEX>0.5 from the left.

We can recognize the characteristic increase of positive compounds in the solid box compared to that in the dotted box. This kind of visualization is effective to understand the nature of the distribution, and to detect nonsense rules coming from the accidental distribution changes.

DISCAS writes an optional RHS terms upon requests. That is, it denotes an attribute value pair if it has a high correlation to the main condition.

```
THEN O = n        34.2%->100.0%; BSS:8.23
THEN C-C-Cl = y 26.6%-> 78.9%; BSS:5.21
```

The absence of O is inferred directly from the absence of HBA. The presence of C-C-Cl is also highly correlated with the main condition, and it may be a point to be considered in further research.

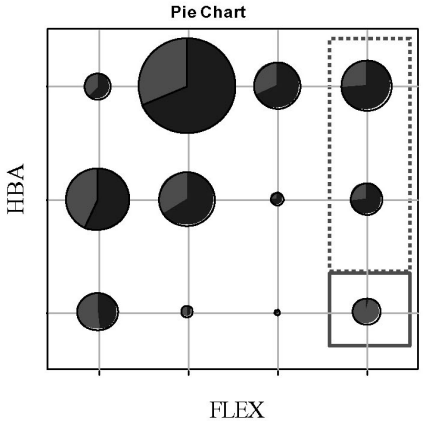


Figure 1. Distribution of pos/neg cases against categorized HBA and FLEX in male mouse data set.

3.3 Prediction of test data

Rules by the cascade model do not intend to give high classification accuracy, but their aim is to give valuable insights for further study. In fact, it is an interesting rule if the positive probability decreases from 0.9 to 0.5, but it does not work in the classification task. However, some rules possess enough accuracy to be used for classification. For example, the rule in the previous section can be interpreted in the following form.

```
IF [HBA = 0] and [FLEX > 0.5]  THEN  [MM = p]
    Cases: 19;  Accuracy: 94.7%;  BSS(root): 6.09
```

Here, *BSS*(root) is the *BSS* value when we employ the root node as the upper node of a rule. We calculated *BSS*(root) values for all rules. Five classification rules with the largest *BSS*(root) values are selected for positive and negative classes, respectively. If no new compounds in the test data set match the rule condition, we select a rule with the next largest *BSS*(root). The sample rule described above has a large *BSS*(root) value, but it could not find applicable compounds in the test data set.

The results of classifications are shown in Table 1. "?" means that the test compound has not matched any of the rules condition. When rules give conflicting classifications, "pos", "equ", or "neg" are assigned depending on the accuracies of the applied rules.

Table 1. Number of classifications in the four test data sets.

	pos	equ	neg	?
MR	18	30	111	26
FR	29	2	61	93
MM	11	9	123	42
FM	38	6	71	70

3.4 Strong rules

Table 2 shows rules with *BSS* >3.0. We could find no such strong rules in the rat data sets. When plural rules share the same main condition in an application to the same data set, weaker rules are shown as comments to the strongest rule. We think that this expression is useful to develop hypotheses for further study. Another interesting information comes from the strength comparison of the same rule among different species. For example, the distribution changes along FM-1 rule are MM: (.40 .60)/228 →(.69 .31)/49, FM: (.42 .58)/227→(.73 .27)/48, MR: (.47 .53)/227→(.61 .39)/44 and FR: (.35 .65)/234→(.48 .52)/44.

Table 2. Strong rules (*BSS*>3.0) derived from four data sets.

No.	Main condition	Preconditions	Changes in distribution †	*BSS*
MM1	[HBA = 0]	[FLEX > 0.5]	(.43 .57) / 79 → (.95 .05) / 19	5.08
	If no precondition is applied, (.38 .62) / 336 → (.65 .35) / 55, *BSS*=4.03. If [C-c:c:c: n] is the precondition, (.40 .60) / 228 → (.69 .31) / 49, *BSS*=4.26. If [c:c-N: n] is the precondition, (.39 .61) / 227 → (.65 .35) / 55, *BSS*=3.79.			
MM2	[C-Cl: y]	[C-O: n]	(.40 .60) / 214 → (.68 .32) / 44	3.56
	The percentage of [HBD = 0] also changes 52%→98%, [c:c-N = n] changes 62%→98%.			
MM3	[Dipole > 4]	[Cl: y]	(.54 .46) / 94 → (.07 .93) / 14	3.11
FM1	[HBA = 0]	[C-c:c:c:c: n]	(.42 .58) / 230 → (.73 .27) / 48	4.54
	The percentage of [c:c : n] also changes 58% → 90%. If [C-c:c:c:c:c:c:c : n] is the precondition, (.42 .58) / 238 → (.73 .27) / 48, *BSS*=4.46. If [C-c:c : n] is the precondition, (.43 .57) / 227 → (.73 .27) / 48, *BSS*=4.37.			
FM2	[N: y]	[c:c:c-N: n], [O: n]	(.57 .43) / 79 → (.16 .84) / 19	3.22
FM3	[HBA: y]	[C-O: n], [c:c : n]	(.57 .43) / 72 → (.16 .84) / 19	3.22
FM4	[C-C-O: y]	[C-Cl: y], [N: n]	(.60 .40) / 52 → (.00 .10) / 9	3.20

† In (*pos neg*)/#*case*, *pos* and *neg* show the probabilities of positive and negative cases, respectively. #*Case* denotes the number of cases. Distributions before and after the application of the main condition are shown.

4 Concluding Remarks

Detailed discussion of the individual rule in Table 2 is beyond the scope of this paper. But, useful insight for the carcinogenic mechanism is expected if we inspect the 2D and 3D structures of the molecules cited in the strong rules.

Lastly, we have to note that the pruning condition is too tight to mine all meaningful rules. Further developments of DISCAS system are necessary to scale up the computation size and to make things easier in the interpretation process of derived rules.

References

[1] Helma, C., King, R.D., Kramer, S., Srinivasan, A.: The Predictive Toxicology Challenge for 2000-2001, http://www.informatik.uni-freiburg.de/~ml/ptc/index.html.

[2] Okada, T.: Finding Discrimination Rules Using the Cascade Model, *J. Jpn. Soc. Artificial Intelligence*, **15**, pp.321-330 (2000).

[3] Gini, C.W.: Variability and Mutability, contribution to the study of statistical distributions and relations, *Studi Economico-Giuridici della R. Universita de Cagliari* (1912). Reviewed in Light, R.J., Margolin, B.H.: An Analysis of Variance for Categorical Data, *J. Amer. Stat. Assoc.* **66,** pp.534-544 (1971).

[4] Okada, T.: Sum of Squares Decomposition for Categorical Data, *Kwansei Gakuin Studies in Computer Science*, **14**, pp.1-6 (1999). http://www.media.kwansei.ac.jp/home/kiyou/kiyou99/ kiyou99-e.html.

[5] Okada, T.: Rule Induction in Cascade Model based on Sum of Squares Decomposition, *Principles of Data Mining and Knowledge Discovery* (*Proc. PKDD'99*), pp.468-475, *LNAI* 1704, Springer-Verlag (1999).

[6] Okada, T.: Efficient Detection of Local Interactions in the Cascade Model, *Proc. PAKDD2000, LNAI*, Springer-Verlag (2000).

[7] Okada, T.: SAR Discovery on the Mutagenicity of Aromatic Nitro Compounds Studied by the Cascade Model", Proc. Int. Workshop KDD Challenge on Real-world Data, pp.47-53, PAKDD-2000 (2000).

分子の構造類似性にもとづくデータマイニング

研究分担者　高橋 由雅　（豊橋技術科学大学工学部）
研究分担者　加藤 博明　（豊橋技術科学大学工学部）

１．背景と目的

「ＡはＢに似ている」、あるいは「ＣはＤと××が似ている」といったいわゆる“類似性”の概念は科学における様々な問題解決の場で利用される極めて重要な概念の一つである。このことは化学の分野においても例外ではなく、その対象となる分子の間の類似性がしばしば言及される[1,2]。特に、これまでの明示的な部分構造マッチングとは別に、“似た構造”あるいは“似た反応”といったいわゆる化学的な“類似性”の概念を如何に取り扱うかは、関連分野におけるコンピュータのより高度な利用を図る上で極めて重要な問題の一つであり、こうした分子の類似性の概念に基礎を置く、より柔らかな構造情報処理に向けた新たな技術の確立が望まれている。化学物質の種々の性質はその化学構造と密接に関連していることは明らかであり、その関係は次のように表わすことができる。

Molecular Property = f (chemical structure)　　(1)

このことは、化学構造が変わればその物質の性質もこれに応じて変わるとの考えを示したものにほかならない。言い替えれば、化合物分子の種々の性質についてその類似性を検討・評価することは、その起因となる化学構造の類似性を検討・評価することと等価な問題と見なすことができる。分子構造の類似性評価に関する研究は新薬開発における候補物質の構造設計や化学物質のリスク評価における分子の特性予測問題に関連して現在活発に研究が進められている分野である。これまで、これらの研究では、化合物間の構造類似性を評価するための構造情報記述子として主に部分構造特徴が用いられてきた。しかしながら、こうしたアプローチではその類似性評価は構造特徴を記述するために予め定義された部分構造集合に大きく依存することが避けられない。そこで、本研究では二次元及び三次元の二つの異なる構造表現レベルで、こうした事前の定義部分構造を必要としない新たな手法とそのデータマイニングへの応用について検討を行なった。また、筆者らは三次元分子構造特徴解析にもとづく知識発見の視点から、種々の生体機能発現に重要なタンパク質のモチーフ構造の自動同定とこれにもとづく三次元モチーフ辞書の作成を進めており、これらの現状についても併せて報告する。

２．検討内容

２．１　TFS を利用した薬物候補構造のデータマイニング

ここでは、まず初めに筆者らが先に提案したトポロジカルフラグメントスペクトル（TFS）[3] による新たな構造情報の記述手法を利用した構造類似性の定量的評価のための方法をもとに、特定の部分構造の有無のみならず化学構造全体の漠然とした類似をも考慮したより柔らかな構造情報の取り扱いと化学データマイニングへの応用に向け、その有用性を実用規模のデータベースを用いて検討した。

＜方法＞

（１）TFS による構造特徴の定量的記述表現

TFS とは化学物質の構造式から可能な部分構造を列挙し、その数値的な特徴づけにもとづいて化学物質のトポロジカルな構造プロフィールを多次元数値ベクトルとして表現しようとするものである。その生成手順は、(1)構造情報の記述表現に際しては化学構造式中の原子を点（頂点）、結合を辺と見なし、原子や結合の種類の相違を区別する重み付きグラフ（化学グラフ）として取り扱う。ただし、水素原子はすべて省略する。(2)与えられた構造式に対応する化学グラフから可能なすべての部分グラフを列挙する。ここでは、親グラフ（もとの化学グラフ）中の異なる要素からなる部分グラフについては同型のものも全て考慮した。(3)次に、得られた個々の部分グラフの定量的特徴づけを行なう。これらの特徴づけには様々な方法が考えられる。たとえば、与えられた親グラフの各頂点原子をその隣接原子の数でラベルづけし、生成された部分グラフをこれらの総和によって特徴づけを行なえば化学構造を単純グラフと見なした場合の骨格のトポロジーを表わす特性スペクトルを得ることができる。また、生成された部分グラフを各部分グラフ中の頂点に対応する原子の質量数の総和（フラグメント重量）によって特徴づければ、原子の種類を考慮した構造フラグメントに関する特性スペクトルを得ることも可能である。このようにして特徴づけられた個々の部分グラフ（構造フラグメント）集合をもとに、その特徴指数に従って度数を調べ、その結果をヒストグラム表示したものがここでの TFS となる。これら、TFS 生成手順の概要を図 1 に示す。TFS は一種のデジタルスペクトルと見なすことができる。これはまた多次元数値パターンベクトルとして取り扱うことができ、そのパターン間の類似性評価に対しては、種々の類似度関数の適用が可能となる。尚、本研究では後者の構成原子の質量数による特徴づけの方法を用いた。また、結合水素原子を有するものについてはそれを含めた拡張原子として各頂点の重みづけを行なった。

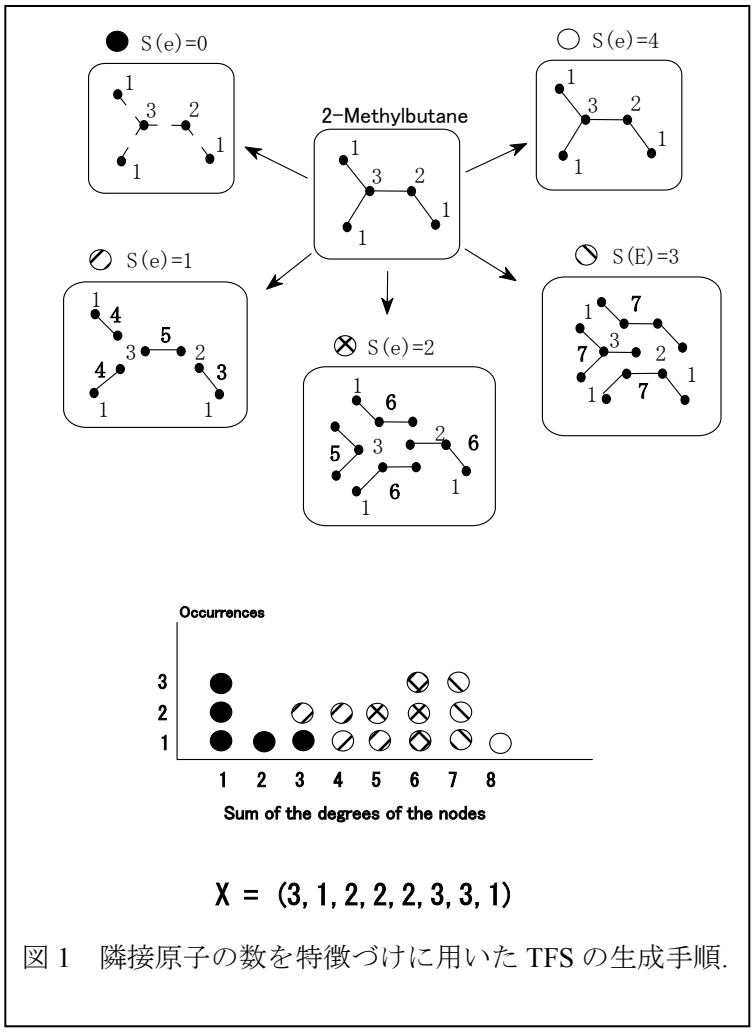


図 1　隣接原子の数を特徴づけに用いた TFS の生成手順.

（２）類似度評価関数

化学物質の構造類似性の評価に際しては、「構造情報をどのように表現するか」がきわめて重要となることは言

うまでもない。この他、もう一つ、その評価結果を大きく左右するものに評価関数を挙げることができる。すなわち、定量的に記述表現された構造情報の数値記述子を「どのような評価関数を用いて評価すればより適切な結果を得ることができるのか」、という問題である。

ここでは、このような視点からいくつかの異なる類似度（あるいは相違度）関数を対象とし、TFS 法を利用した構造類似性評価における各々の関数の適否並びに相互の結果の比較検討を行なった。類似度の評価には様々な評価尺度の利用が考えられるが、本研究では、ユークリッド距離（S_{ED}）、Tanimoto 係数（連続値データ）（T_C）、Tanimoto 係数（バイナリ・データ）（T_B）、Cosine 係数（S_C）、ピアソンの積率相関係数（S_P）の五つの類似度（相違度）関数について検討を行なった。

(a) ユークリッド距離(S_{ED})：
$$S_{ED} = \sqrt{\sum (x_{ik} - x_{jk})^2} \tag{2}$$

(b) Tanimoto 係数(T_C)：
$$T_C = \frac{\sum (x_{ik} x_{jk})}{\sum x_{ik}^2 + \sum x_{jk}^2 - \sum (x_{ik} x_{jk})} \tag{3}$$

(c) Tanimoto 係数（二値データ）(T_B)：
$$T_B = \frac{C}{Q + D - C} \tag{4}$$

(d) Cosine 係数(S_C)：
$$S_C = \frac{\sum (x_{ik} x_{jk})}{\sqrt{\sum (x_{ik})^2 \sum (x_{jk})^2}} \tag{5}$$

(e) Pearson's 相関係数(S_P)：
$$S_P = \frac{\sum (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum (x_{ik} - \bar{x}_i)^2 \sum (x_{jk} - \bar{x}_j)^2}} \tag{6}$$

ここで、$x_{ik,}, x_{jk}$ はそれぞれ化合物 i および j の k 番目の記述子の値を表わす。尚、式(4)においては各 TFS をピークの有無にしたがって１または０の値をとる２値スペクトルに変換している。ここでは、Q は Query 構造の TFS 中で値が‘１’の要素の数、D はデータベース中の比較する構造の TFS 中に含まれる値が‘１’の TFS 要素の数、C は Query 構造およびデータベース構造の両者にともに値‘１’をもつ要素の数を表わす。

（３）データ・セット

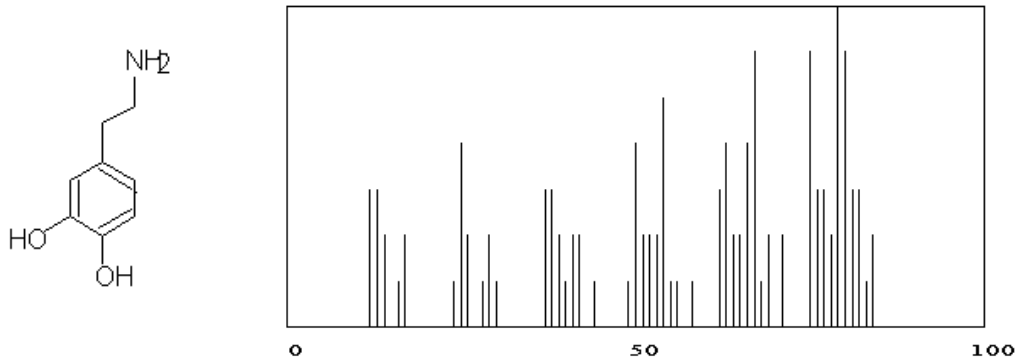
本研究では World Drug Index（WDI）より、ドーパミン活性を有する化合物を抽出し、各々ランダムに抽出した化合物を加えることによって薬物構造データベースを作成し、テストデータベースとして用いた。作成したデータベースは 3,609 化合物からなる。

＜結果及び考察＞

ここでは上記の考えをもとに、化学物質の構造類似性評価における TFS 法の有用性と薬物構造データベースを対象としたデータマイニングへの応用の可能性を検証するため、薬物構造データベース中の

全ての化合物に対する TFS を生成したデータベースを作成し計算機実験を行なった。本実験では、TFS の生成に際してはサイズ（生成部分構造に含まれる結合の数）が５までのフラグメントを用いた。これらを利用し、種々の化合物構造を Query とした類似構造検索を試みた。

ドーパミンを Query とし、評価関数に単純ユークリッド距離を用いて構造類似性検索を試みた。Query として用いたドーパミンの化学構造とそのサイズ５までの生成フラグメントにもとづく TFS を図 2 に



ドーパミン　　ドーパミン TFS 部分スペクトル

図 2　ドーパミンの化学構造式と生成部分構造を質量数で特徴付けた TFS.

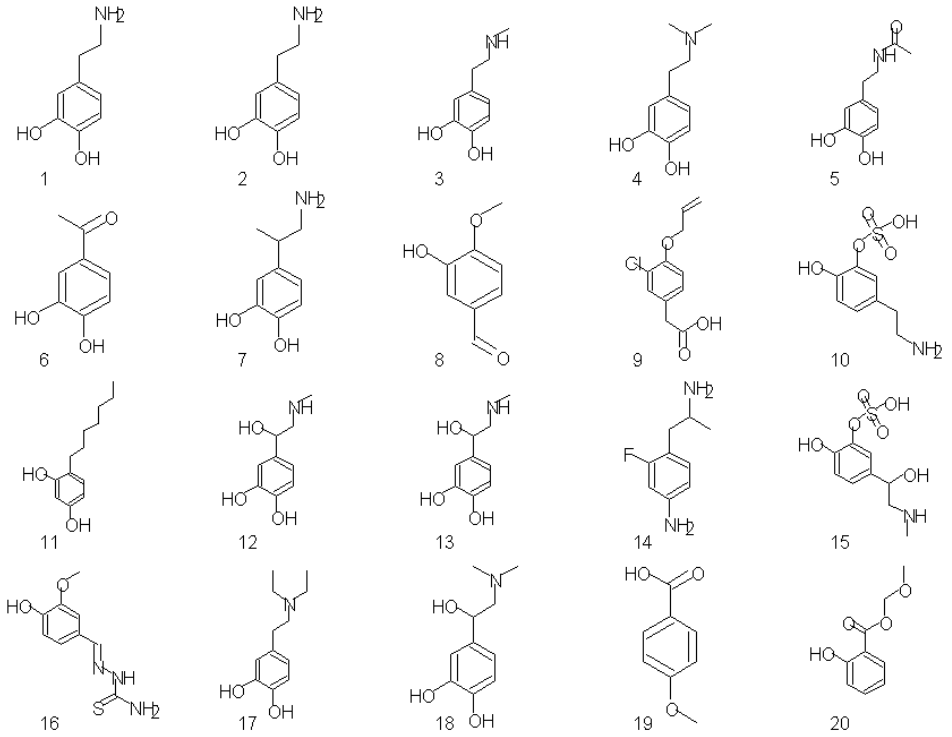


図 3　TFS 法によるドーパミンを Query とした場合の類似構造検索の結果.
（ユークリッド距離を評価尺度として用いた場合の上位 20 化合物の化学構造）

示す。また、構造類似性検索にもとづくこの場合の上位 20 化合物の化学構造式の一覧を図 3 に示す。図 3 に示されるように、直感的にも構造的によく似ているものが上位にランク付けされていることが分かる。ここでは Query であるドーパミン自身がデータベース中に異なる商品名で 2 件含まれており、これらの二つが距離ゼロで最初に検索されていることが分かる。

次に、TFS 法における類似性評価に際しての種々の類似度関数の適否について検討を行なった。評価尺度としてユークリッド距離（S_{ED}）のほか、方法の部で述べた Tanimoto 係数（T_C）、Tanimoto 係数（バイナリ）（T_B）、Cosine 係数（S_C）、ピアソンの積率相関係数（S_P）について検討を行なった。これらの結果をまとめて表 1 に示す。表 1 のユークリッド距離と連続型変数を対象とした Tanimoto 係数（T_C）を用いた場合の結果に注目すると、上位 20 化合物で 9 割以上が共通の構造をヒットしていることが分かる。これは他の Query 構造を用いた場合もほぼ同様であった。このことは TFS を基礎とした

表1　ドーパミンをQueryとした種々の類似度関数による類似性解析の結果.

Rank	No.	ID	Value	Rank				Act.
			S_{ED}	T_B	T_C	S_C	S_P	
1	549	DOPAMIHCL	0.000	1	1	1	1	1
2	553	DOPAMINE	0.000	2	2	2	2	1
3	538	DEOXYADRE	5.385	14	3	4	4	0
4	539	DIMEDOPNN	6.325	-	4	3	3	1
5	557	ACDOPAMIN	7.348	-	5	7	9	0
6	961	DHOACET34	8.544	-	7	14	14	0
7	544	MEDOPAMIB	8.602	-	6	5	5	1
8	962	ISOVANILL	9.695	-	10	-	-	0
9	3456	ALCLOFOLA	9.849	-	13	-	-	0
10	2232	DOPAMSUL3	9.899	-	16	-	-	0
11	1477	HEPTYLRES	10.050	-	11	18	20	0
12	542	RACEPINEP	10.488	-	8	8	7	0
13	2238	ADRENAASC	10.488	-	9	9	8	0
14	1039	NBF027	10.536	-	19	-	-	0
15	2231	ADRENSUL3	10.583	-	12	12	13	0
16	3411	VANITHIOZ	10.724	-	15	19	-	0
17	547	DIETDOPAM	10.909	-	17	-	-	1
18	541	METHYLADR	10.954	-	14	10	10	0
19	957	ANISATEPA	11.000	-	-	-	-	0
20	965	SALICYMME	11.180	-	-	-	-	0

構造類似性検索においてはこれらの評価関数はほぼ同様な結果を与えることを示唆している。一方、TFS の２値表現を基礎とした Tanimoto 係数（T_B）は他の類似度関数を用いた場合に比べて大きく異なる検索結果を与えていることが分かる。この場合、先のユークリッド距離を用いた場合の検索結果との重なりは僅か 1 割程度であった。このことは、類似度関数 T_B はスペクトルの強度（頻度）情報を利用していないことによるものと考えることができる。

また、ここでの構造類似性検索実験に Query として用いたドーパミンは神経伝達物質としての作用を示すことが知られている。このことから、上記の検索結果をもとに得られた構造類似化合物についてのドーパミン活性の有無を WDI データベースを用いて調べてみたところ、複数の化合物が同活性を有するものであることが明らかとなった。これらの結果についても合わせて表 1 に示した。Query 以外にもドーパミン活性を有する化合物が 3 件含まれていることが分かる。

２．２　タンパク質三次元構造特徴の探索

ヒトゲノム計画の進展、並びにタンパク質構造決定技術の進歩に伴い立体構造のデータは急速に増加しており、その構造データベースはタンパク質の構造と機能との関係解明など分子生物学上の新たな知識獲得のための基本要素としてその重要性はますます高まっている[4,5]。しかし、タンパク質構造の巨大さや複雑さ、さらには近年の急激なデータ数の増大から、手動によるモチーフの検索やその特徴解析はほとんど不可能となっている。そのため、これらのデータベースを有効に活用し、三次元構造特徴の系統的な解析を行なうための方法論の確立、並びに有効なコンピュータツールの開発が切望されている。

このため筆者らはこれまでに、グラフ論的な部分構造検索技法を基礎とした三次元モチーフ構造検索アルゴリズム、さらには質問構造の設定を要求しない複数タンパク質間の三次元共通構造特徴（新規モチーフ候補部位）の自動認識に向けての基本アルゴリズム並びに対応するシステムの開発を進めてきた[6,7]。一方、アミノ酸配列レベルでの特徴解析は比較的古くからなされており、このうち Bairoch は配列レベルのモチーフ情報を広く収集して電子化したモチーフ辞書 PROSITE を作成、公開している[8]。本研究では、この既存の知識ベースともいえる PROSITE に登録されている配列モチーフパターンに注目し、これに対応する三次元部分構造情報を網羅的に集積するためのプログラムの開発を試みた。また、集積した対応三次元部分構造群を定量的に比較・分類し、各配列モチーフに対応する代表三次元幾何パターンを決定するための方法についても併せて検討した。

表 2　PROSITE 中の配列モチーフの例.

モチーフ名	パターン
Kringle	[FY]-C-R-N-P-[DNR].
Zinc finger C2H2	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.
EF-hand	D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW].

* パターン中で、[ ] 内はその中のどれかのアミノ酸と、{ } 内はその中以外のアミノ酸と、x(n,m)は n 個から m 個の任意のアミノ酸とマッチすることをそれぞれ意味する.

＜方法＞

（１）配列モチーフ検索

PROSITE では、配列モチーフを一種の正規表現を用いてパターンとして定義している（表 2）。本研究では Protein Data Bank (PDB [9]) に登録された三次元構造既知のタンパク質構造情報を対象に、PROSITE で定義された配列モチーフ部位を検索し、その対応する三次元部分構造をデータベースファイルへ集積する。ここでは、タンパク質の三次元構造情報はその構成アミノ酸残基を単位として取り扱い、各アミノ酸のアルファ炭素(Cα)原子の座標で代表して縮約表現することとした。また処理の効率化のために、あらかじめ PDB ファイルから鎖単位でアミノ酸配列情報のみを抽出したファイルと、その三次元構造情報（結合表形式）を抽出したファイルの二つを作成し利用した。

本研究では、ユーザの目的によって対話的あるいはバッチ的な使い分けできるように、次の四つの実行モードを定義し、Microsoft Windows 上で VisualBasic 6.0 を用いて検索プログラムの実装を行なった。プログラムの実行画面例を図 4 に示す。

① ある一つのタンパク質の配列データを対象に、ある一つの配列モチーフを指定し検索するモード (Single-Single)

② データベースに保存された複数のタンパク質の配列データを対象に、ある一つの配列モチーフを検索するモード (Multi-Single)

③ 一つの配列データを対象に、複数の配列モチーフを検索するモード (Single-Multi)

④ 複数の配列データを対象に、複数の配列モチーフを検索するモード (Multi-Multi)

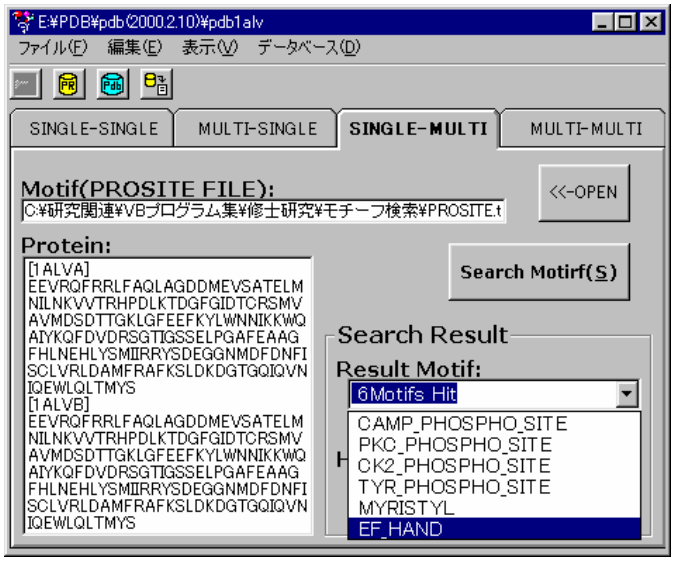


図 4　配列モチーフ検索プログラムの実行画面例.

バッチモードでの検索でヒットしたタンパク質及び対応部位情報は一時ファイル（MS-Access 形式のデータベースファイル）に登録され、必要に応じてここから対応部位の一覧情報を記述したファイル（インデックスファイル）や、各対応部分構造の結合表ファイルを取得することができる。ここで、インデックスファイルには一行目にヘッダ情報としてモチーフの PROSITE 中での ID 番号、対応部分構造数、そのサイズ（構成残基数）の種類とその内訳を、二行目以降は対応部分構造について、シーケンシャル番号、PDB コード（必要に応じて鎖名を付加）、サイズ、開始及び終了位置番号を記述した（表 3）。

また、得られた三次元部分構造をグラフィックス表示するためのインターフェースプログラムも併せて実装し、結果を視覚的に検証できるよう工夫した。表示には対応する PDB ファイルを参照し、表示ツールとして MDL 社の Chemscape Chime プラグインを、対応部位の指定には RasMol スクリプト形式を用いた。作成したプログラムでは、前述のインデックスファイルを入力し、表示モデル等の指定を行ない、HTML ファイルの出力を行なう。表示画面例を図 5 に、ユーザが指定可能なメニューの一覧を以下に示す。

① 表示する三次元部分構造の指定
② 表示モードの指定
（全構造、部分構造、両方）
③ 表示モデルの指定
（Ball&Stick、Spacefill、Ribbons 等）
④ 画面のサイズと分割数

表 3 インデックスファイルの例.

```
PS00205        14   2   9  10
0 1A8E      9 --- Y   96 - D  104
1 1A8E     10 --- Y   95 - D  104
                  .
                  .
12 1CE2A   10 --- Y   92 - G  101
13 1CE2A   10 --- Y  433 - A  442
```

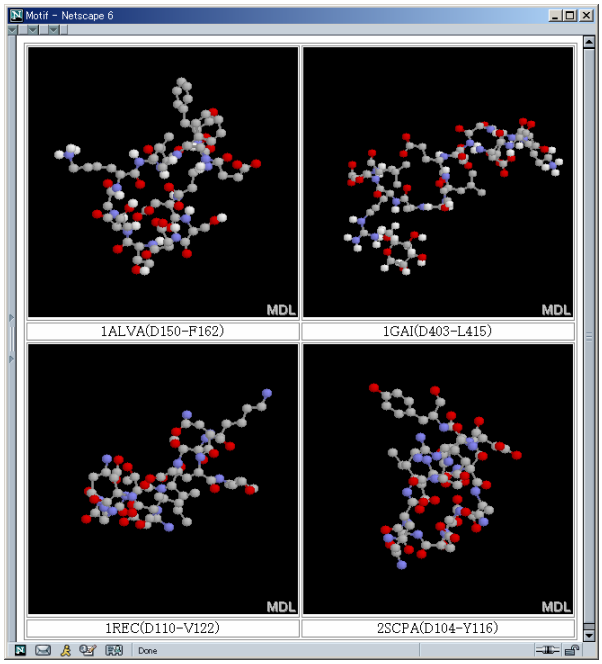


図 5 対応三次元部分構造のグラフィックス表示例.

（2）構造クラス分類

以上で集積された共通の配列パターンを持つ三次元部分構造を検証した結果、その構造が大きく異なるものがいくつか存在した。そこで、次にこれらの三次元部分構造を定量的に比較・分類する方法について検討した。前述のとおり、本研究では三次元部分構造をアミノ酸残基単位で取り扱い、それぞれ Cα原子の座標で代表して表現する。このようにして表現した二つの部分構造内の対応する 2 点（Cα原子）間の距離の差（絶対値）の平均値を求め、これをこれら二つの構造間の類似度（相違度）の尺度と定義した（式 7）。なお、PROSITE の柔軟なパターン定義により、サイズ（構成アミノ酸残基数）が異なる部分構造が発生した場合は、より小さい方の部分構造を基準とし、配列順序を維持したまま可能な組み合わせを全て列挙する。そして、それぞれ上記の相違度の値を計算し、そのうちの最小値をこれら二つの構造の相違度と定義した。

$$\text{相違度}(A,B)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sqrt{(d_A(i,j)-d_B(i,j))^2}}{n^2} \qquad (7)$$

ここで、n は部分構造のサイズ（アミノ酸残基数）、$d_A(i,j)$, $d_B(i,j)$はそれぞれ部分構造 A, B の i 番目の残基と j 番目の残基の間の距離を表わす.

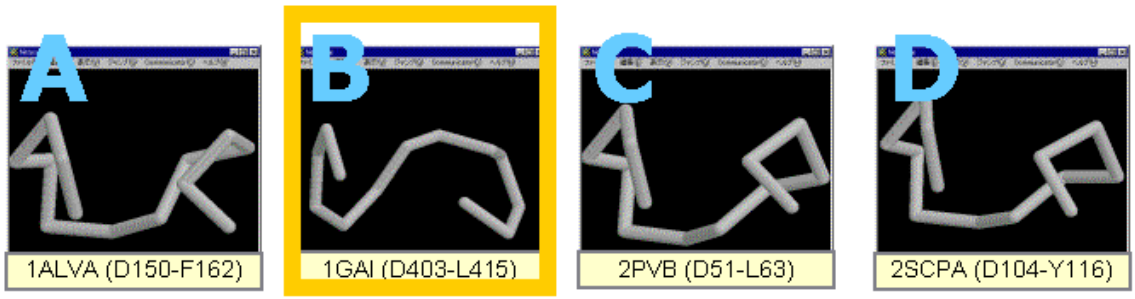


図 6　比較する四つの三次元部分構造.
（EF-hand モチーフ対応部位の一例）

表 4　相違度行列による構造クラス分類結果例.

	A	B	C	D
A	0	811	101	62
B	811	0	802	786
C	101	802	0	56
D	62	786	56	0

＊ クラス(1)：A, C, D、クラス(2)：B

ある一群のタンパク質部分構造（共通の配列パターンを持つもの）を対象に、全てのペア間の相違度を求め、相違度行列を生成する。例えば、表 4 は図 6 に示す四つの部分構造に対する相違度行列の例である。ここである閾値（例えば 100）を設定すれば、その閾値以内の構造同士は同じクラスに属し、それより大きいもの同士は別々のクラスに属するものとみなすことができる。閾値の与え方により結果として得られる構造クラスの数やサイズは異なるが、適切なクラス分けとなるような閾値をあらかじめ決定しておくことは難しい。そこで本研究では、ある指定した範囲内で閾値を変化させ、その結果得られるクラスの情報を一覧表示するユーザインターフェース・プログラムを実装し、対話的にクラス分類を行なえるよう工夫した。

（3）代表幾何パターンの決定

以上によりいくつかの構造クラスが得られたら、次に各クラスを代表する一つの三次元幾何パターンを決定する。本研究では、各部分構造について、当該クラス中の他の全ての部分構造との相違度の和を算出し、その合計値が最小となるものをそのクラスの代表パターン（すなわち、クラスの情報を反映する最も平均的な三次元構造）であると定義した。表 4 の例の場合、A・C・D の三つの部分構造が属するクラス(1)の代表パターンは D、クラス(2)については B（自明）と決定される。

＜結果および考察＞

PDB (Rel. #89)から抽出したタンパク質（902 鎖）テストデータセットに対して、PROSITE (Rel. #16.0)中の全パターン（ただし、N-, C-両末端の終端指定子を含む特殊なパターンは除外）1,299 件を対象に Multi-Multi モードによる三次元構造モチーフデータベースの作成実験を試みた。その結果、少なくとも一つの三次元部

分構造がヒットした配列モチーフは 464 件であった。次に、このうちヒットした部分構造数が 3〜50 のものについてクラス分類-代表パターン選出実験を引き続き行なった。

このうち、例えば EF-hand モチーフ(ID PS00018)に対応する三次元部分構造は延べ 12 個（一つのタンパク質構造中に複数のモチーフ部位がヒットするものも含む）検索され、構造クラス分類を行なうと、閾値の変化により１〜４のクラスが得られた。このうち、３クラスに分割されたケースについてみると、第１のクラスには 10 個の部分構造が属し、その代表パターンは 2SCPA (D104-Y116)と決定された。一方、他の二つのクラスはいずれも属する部分構造が一つのものであった。これら（1GAI(D403-L415) 及び 1GOH(D75-V87)）について調べると、そのタンパク質の機能や周辺環境（カルシウムイオンや、E, F と呼ばれる前後の二つのヘリックスの存在 [5,10]）から、本モチーフに該当しないノイズ成分であったことが確認できた。

３．まとめと今後の課題

以上、本研究では実用規模の薬物データベースを用いた類似構造検索の実験を通じて、TFS 表現にもとづく構造類似性評価を基礎とした化学データマイニングの有用性を示した。しかしながら、TFS を利用した構造類似性解析においてはどの類似度関数が最適であるかを結論づけることは困難であり、現状では異なる視点から定められた複数の類似度（相違度）関数を目的に応じて利用することが重要と思われる。今後の課題としては TFS を利用した化合物の化学クラスや活性クラス分類への応用、並びに TFS の各ピークの意味を解析するためのツールの開発、さらにはこれらを利用した知識発見への応用などが考えられる。

一方、タンパク質の三次元構造特徴探索では PROSITE 中で定義された配列モチーフに対応する三次元部分構造の集積・分類を行なった。特に、EF-hand モチーフの例では構造クラス分類の結果、ノイズ成分を除去した主要クラスが生成され、その代表幾何パターンを得ることができた。今後は、得られた部分構造（特に代表パターン）を基礎とし、別途開発した三次元モチーフ検索プログラムを用いてタンパク質構造データマイニングのための三次元構造特徴辞書の作成を進めるとともに計算機実験を通じてその有用性を明らかにしたい。

[参考文献]

[1] M.A.Johnson and G.M.Maggiora: *"Concepts and Applications of Molecular Similarity"*, Wiley, New York, (1990).

[2] R.Carbo: *"Molecular Similarity and Reactivity",* Kluwer Academic Publishers*,* Boston (1995).

[3] Y.Takahashi, H.Ohoka and Y.Ishiyama: "Structural Similarity Analysis Based on Topological Fragment Spectra", *Advances in Molecular Similarity,* Vol.2, 93-104 (1998).

[4] 金久實: "*ヒューマンゲノム計画* ", 共立出版 (1997).

[5] C.Branden and J.Tooze: "*Introduction to Protein Structure"*, Garland Publishing, New York (1991).

[6] H.Kato and Y.Takahashi: "Three-Dimensional Structural Feature Search of Proteins", *Bull. Chem. Soc. Jpn.*, **70**, 1523-1529 (1997).

[7] H.Kato and Y.Takahashi: "Automated Identification of Three-Dimensional Common Structural Features of Proteins", *J. Chem. Software*, **7**, 161-170 (2001).

[8] A.Bairoch: “PROSITE: a Dictionary of Sites and Patterns in Proteins”, *Nucleic Acids Res.*, **19**, 2241-2245 (1991).

[9] F.C.Bernstein, *et al.*: “The Protein Data Bank: a Computer-based Archival File for Macromolecular Structures”, *J. Mol. Biol*., **112**, 535-542 (1977).

[10] R.H.Kretsinger : “Structure and Evolution of Calcium-Modulated Proteins”, *CRC Crit. Rev. Biochem.*, **8**, 119-174 (1980).

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

研究代表者　大澤幸生　（筑波大学ビジネス科学研究科）
研究分担者　寺野隆雄　（筑波大学ビジネス科学研究科）

背景と目的

「知識が有用である」とは、利用者にとって理解が容易であり目的に応じて的確に使用できること、また、利用者の創造性を刺激しうる機能を備えることを意味する。そのために、知識の需給関係に注目し、知識の候補を供給するシステムとそれを解釈・選択・利用する専門家とのインタラクションを通じて、知識を評価・選択できるような方式を確立するのがわれわれの目的である。この方式は、利用者個人あるいはグループの主観までを評価尺度に含み、従来研究されてきた客観的な基準での知識評価方法を超えるものとする。

a. 客観的意思決定の支援：医学界において Evidence-Based Medicine（EBM）の考え方が広まりつつある[1][2]。この EBM は，病気の診断・治療に科学的な根拠を求め、確実な医療の実現を目指すものである。臨床家個人の経験や権威者の意見が尊重されてきた従来の医療を見直し、医学論文等に示された科学的根拠に基づき、標準的な医療を提供しようというものである。医療が高度化･複雑化した現在、医療過誤の防止や医療費抑制の面からも医療の標準化に期待が寄せられている。EBM は３つの段階によって実現される。それは、（１）Evidence を作成し、（２）Evidence を利用可能にし、（３）Evidence を利用するという段階である。各段階において情報技術による支援が不可欠であり、データ解析、データベース構築、検索処理を主体とした情報処理が求められる。

EBM が従来のデータマイニングあるいはデータベースからの知識発見のタスクと異なるのは、i)はじめに大量のデータが与えられるわけではなく Evidence 作成に専門家が積極的に関与しなければならないこと、ii)ならびに各プロセスにおいて得られた知識・情報を医学面からもデータマイニングの観点からも、インタラクティブに評価しなければならない点である。この意味において、EBM とデータマイニング、知識の評価の問題は、アクティブマイニングのかっこうの研究テーマである。

Evidence-Based Medicine について　EBM の実践は、系統的な研究や臨床疫学研究などの外部から提供される根拠と、臨床家個人の臨床的専門能力を統合して、診療に望むことを意味する[3][4]。その具体的手順は下記のように示される。

１）患者の問題の定式化
２）能率的な情報の収集
３）情報の批判的吟味
４）患者への情報の適応
５）事後評価

情報を Evidence として利用する際には、その情報の質を厳格に評価する必要がある。Evidence の質は、医学研究のデザインと利用できる統計的手段・データマイニング手法に

大きく依存している。Evidence の水準に関しては、いくつかの機関から示されており、Table １に示す Evidence の水準は米国健康政策・研究局のものである。ここでは、最も強力な Evidence として、複数のランダム化比較試験の結果をメタアナリシスで統合したものが挙げられている。続いて、単独のランダム化比較試験の結果、ランダム化されていない比較試験の結果、準実験的研究の結果といずれも実験的研究が上位にあり、ケースコントロール研究など観察的研究の結果、専門委員会・権威者の意見や臨床経験は下位に格付けされている。実験的研究による効果の証明が重視される姿勢は明らかである。しかし、EBM の取り扱う問題は、薬効のような治療効果に関するもの以外に、診断やリスク・予後の評価など診療全般に関わっており、例えば、有害物質のリスク評価など倫理的に実験が不可能な場合があり、観察的研究の必要性は否定できない。

一方、EBM 実践のための情報を利用しやすい環境で提供する電子図書館の構築が進められている。コクラン共同計画[5]では、専門家による評価済み Evidence のデータベース化が進められている。その他、米国政府機関である AHRQ（Agency for Healthcare Research and Quality）が作成した EBM 準拠のガイドラインや関連文書が掲載された NGC（National Guideline Clearinghouse）[6]や、専門家が統制語によるインデックスをつけた論文が収録された PubMed[7]などのサービスがある。

b．主観的発見の支援： 欧米を起点として客観的医療に傾倒しようという方針に，わが国の医療政策は向かおうとしている。しかし一方，これとは別個に、やはり医師の持つ勘や判断の個人的特性に立ち戻ろうという考えは根強い。それらは場合によって主観的な判断といわれ，現時点で万人が良いと認めると限らないという意味で，マスコミなどの大衆意見としては否定的に評価されることがあるが，やはり新しい病気や難病についての判断において，未定着の医療技術は誰かが治療の鍵を発見してこれを広めるというような，人によって判断の違う段階が必要である。

今ではデータから一般的法則を得るための学問と見られるようになった統計学においても，初期にはピアソン（Pearson）が「自然が法則に従うのではなく，人が自然に法則を与えるのだ」と述べて，法則の客観性よりも人の主観の産物を重視した[21]．統計学（Statistical Science）そのものもピアソンの師ゴルトン（Galton）によって，人がグループで議論する対象を膨大なデータから得る方法と位置付けられ，あくまでも人が法則発見の主人公であった．しかしその後は，さまざまなデータと数学者の登場とともにさまざまな統計理論が打ち出され，統一的な観点で多くのことがらを説明する普遍性志向の強い西欧科学の伝統にしたがって理論は深められていった．このような観点の統一性を客観性と呼び，これに反するとみなされた個人の主観は，むしろ排除される傾向が発生した．

近年，この傾向を見直し，いかに人の主観を意思決定に利用するかという問題を問い直す方向が統計学にも出ている．これは，マーケティングにおける顧客データ分析に代表される実応用へと統計解析が積極的に導入されるに伴い，却って人々のニーズが多様な主観に由来していると認識されたためである．チャンス発見の成果[22]によれば，それらのニーズを理解するためには，データが扱う顧客たちの行動の多様な文脈を把握できる多様な人々の解釈を統合する考え方が有効である．このようなわけで，統計学が新しい時代の創

造と維持に貢献しようとする意識に「チャンス発見」研究の成果が合致した。一見伝統的な統計学の国際シンポジウムにチャンス発見学の動向についての講演が何度か招かれたことは，学術史の流れとして理解しやすい動向である[23].

チャンス発見では，新たな事象が発生したときに、その事象が人の意思決定（この例では医療行為の選択）にもたらす意義を人に先駆けて理解するというプロセスの支援手法を研究している。ここでのポイントは,「人に先駆ける」主観的気付き先行し、これが客観的知識として蒸着するようにすることである。この意味で，このプロセスは，a.の客観的意思決定の支援を行うまでのプロセスに位置付けることができる。チャンス発見学については、われわれは世界をリードして強力に推進し，現在 AAAI の秋のシンポジウムに採用されるなどの経緯をたどっている。

本稿では、a，b の両側面についてアクティブマイニングへのわれわれのチームの貢献の方針と展開を述べる。

検討内容

a．客観的意思決定の支援についての検討と現状

EBM を実現するには、幅広い Evidence の蓄積と EBM 実践の方法論の開発が必要である。本章でははじめに、Evidence 作成におけるデータマイニングの位置付けを明らかにし、その上で応用可能性について述べる。次に、研究結果を統合しより強固な Evidence を作成するメタアナリシス[8]に触れ、分散/並列データマイニングの要素技術であるメタラーニング[9]の応用可能性について述べる。最後に EBM の実践に寄与する支援技術としてエージェント技術への期待について簡単に触れる。

Evidence の作成とデータマイニング技術　Table1 で示されているように、EBM では観察的研究の結果は低い格付けにあり、Evidence としての利用に耐えられない。丹後は、観察的研究の結果が正しい可能性は 10%にも満たないと指摘している[10]。この認識のもとで観察的研究の役割を明確にし、データマイニング技術の医療応用に結びつける必要がある。

臨床試験のような実験的研究には、倫理的、経済的に大きな制約がある。危険因子の評価を扱う疫学調査では暴露実験は実施不可能であるし、経済的な制約から実験的研究に供される命題はかなり絞られたものになる。データ収集の容易さとそのコストの面では、観察的研究が優位であり、また、倫理的問題を含む研究では観察的研究のみが選択肢となる。

観察的研究は下記のように分類される[11]。

・ケース-シリーズ研究
　少数の患者で観察された事項を記述
・ケース-コントロール研究（Retrospective）
　症例と対照の２群について、過去の原因や危険因子を調査
・クロス-セクショナル研究　（Prevalence）
　ある一群の対象者に一時点で何が起きているかを調査・ある疾患や症状の有無を短期間に調査
・コホート研究（Prospective）
　研究対象をある危険因子の有無で分け、一定期間観察し、両群で何が起こるかを調査

上記の分類の中で特に、データ収集の時間的流れの違いによる、前向き研究（Prospective Study）と後ろ向き研究（Retrospective Study）の区別が重要である。通常、データマイニングの立場は、後ろ向きなケース-コントロール研究あるいは現時点でのクロス-セクショナル研究もしくはケース-シリーズ研究である。これらの研究から得られた知識は、交絡因子・バイアスの調整が困難であるとの理由から仮説の域を脱しない。仮説を Evidence として磨きをかけるためには仮説検証のプロセスが必要である。津本が指摘するように、追加的に前向き調査を実施[12]するなり、可能であるならば実験的研究にて検証することが求められる。これはデータマイニングの医療適用において強調されるべき課題である。

従来、観察的研究であろうと実験的研究であろうと、研究者が想定した仮説を検証する立場で統計学的手法が駆使されてきたため、効果があるかないかといった簡単な命題・知識構造が主に扱われてきた。そこで、より複雑な知識構造を探るためにデータマイニング技術を利用することは有用である。

生体の特徴は、ホメオスタシスとゆらぎの２面を持っている点であり、そこから得られる情報にはあいまいさが存在する。例え測定が正確であっても、真の姿を捉らえることは困難である。病的な状態では特定のパターンが出現することで、その診断が可能となるものの、病態の成因もまた多くの要因が関連し因果関係の特定も困難な場合も多く、臨床検査データの判読は容易ではない。今日の医学教科書は、要素還元的な記述にとどまっており、複雑系として捕らえた病態モデルとしての理解が必要であると指摘できる。データマイニングによる医学データ解析への接近は、この契機となる可能性が高い。

ここで、臨床検査分野におけるデータマイニング事例を取り上げる。

稲田らは、血液中の複数の酵素データを用いて病態モデルを構築した[13]。このモデルは、臓器の状態が検査データに反映されるという因果関係を、臨床検査の実データから因子分析で解析・モデル化し、因子スコアの視覚的表現を試みたものである。視覚化によって、臨床家の検査データ評価に関わる負担を軽減し、さらに、重要な病的異常値の見逃しを防止する効果が期待される。

片岡らは、血清蛋白分画データの病態パターンのクラスタリングを行った[14]。蛋白分画データは病態によって特徴のあるパターンが現れることが知られている。自己組織化マップを用いて蛋白分画データの挙動をクラスタリングした上で、得られたクラスターを病態と対応付けるために、分類クラスと他の臨床検査データを属性として決定木分析を行っている。意味付けされた分類クラスは、臨床診断につながる付加価値情報として有用性がある。

ここに示した事例は、新規性や意外性のある知識発見を目指したものではない。ビジネス界でのデータマイニング応用では、競争優位を築く上で、新規性や意外性のある知識が求められ、各種データマイニング手法の開発もこの方向で進んでいるように思われる。しかし、医療応用においては、必ずしも新規性や意外性だけが求められるわけではない。大量データを解析し、既存の医学知識を再評価することや、複雑な知識構造を明らかにするなどの地道で確実な知識発見への要求がある。EBM の実現にはこうした確実な知識ベースの構築が急務である。

確実な知識を得るためには、知識の解釈・評価のプロセスが重要である。例えば、ニュ

ーラルネット等を用いた自動診断システムの開発事例[15]があるが、得られた知識の解釈が極めて困難であり、分類精度の評価だけでは、倫理的な面で実用化は困難であろう。前述の片岡らの研究では、自己組織化マップにより得られたクラスターに、決定木分析を組み合わせて臨床的な意味付けが行われている。知識の解釈のために、データマイニング手法をさらに組み合わせることは有用であろう。

一方、知識評価のプロセスに焦点を当てた手法も開発されている。知識評価のプロセスは必然的に意思決定の問題を伴う。この意味でデータマイニングでしばしば用いられる。Precision/Recall による知識の評価尺度に加えて、知識に関する選好や Utility による評価も重要である。そこには当然、専門家の主観的判断が含まれる。実際、知識発見のタスクにおいては、「あたりまえ」と「無意味」の境界上に「有用な知識」が存在することは良く知られている。

このような観点から、石野らが開発したアンケートデータ分析ツール SIBILE[16]は、対話的進化計算と帰納学習を組み合わせて、解析者が得られた知識を評価しながら対話的にデータマイニングを進めることを可能にしている。筆者らの経験では、対話的操作の中で、既存の背景知識による評価を行うことで過学習が回避される効果も認められた[17]。

Table 1: Evidence の質の分類 (米国健康政策・研究局)

Ⅰa	複数のランダム化比較試験のメタアナリシスによる
Ⅰb	少なくとも一つのランダム化比較試験による
Ⅱa	少なくとも一つのよくデザインされた非ランダム化比較試験による
Ⅱb	少なくとも一つの他のタイプのよくデザインされた準実験的研究による
Ⅲ	比較研究や相関研究、ケースコントロール研究など、よくデザインされた非実験的記述的研究による
Ⅳ	専門家委員会の報告や意見、あるいは権威者の臨床経験

多くのデータマイニングアルゴリズムは、分類精度が最小となるように設計されている。しかし、必ずしも分類精度最小の知識が専門家に支持されるわけではない。臨床検査医学の分野では、検査法の有効性を評価する指標として、感度（疾患のある群の中で陽性を示す割合）と特異度（疾患のない群の中で陰性を示す割合）がある[18]。これらの指標は臨床的判断値を設定する際にも用いられ、トレードオフの関係にある両指標のバランスを考慮することが求められている。筆者らはこれらの指標を知識の評価にとり入れた医療向け SIBILE の開発を行っている。

メタアナリシスとメタラーニングシステム　前節では、Evidence の作成として、生データに対して行われる一次的分析について、データマイニング技術の応用可能性を述べた。ここでは一次的分析を総括する二次的分析についてデータマイニング技術の応用を検討する。

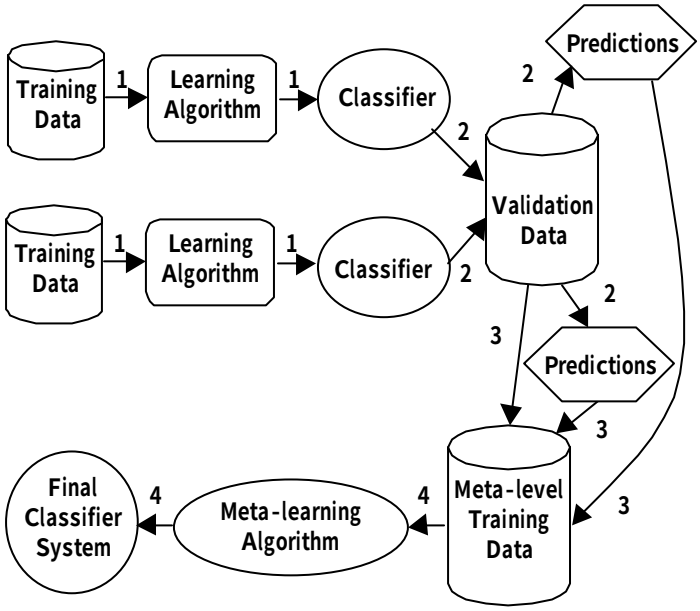


Fig. 1：Meta-Learning の概要

1995 年の Science の論説[19]では、次々と発表されては否定される疫学研究について問題が提起され、疫学者・医学統計学者に対するインタビューによりその原因が探られた。原因として、過大な報道が挙げられる一方、不十分なランダム化のために交絡因子の調整が不十分なことに起因して、バイアスの大きさが各研究で異なる点が指摘されている。この指摘の通り、個々の研究にバイアスが存在することは否定できないことから、そのリスクを回避するために、30 以上の類似研究をまとめて評価することが求められる。ここで、複数の独立した類似の研究結果を分析する手法としてメタアナリシスが用いられる。

メタアナリシス[8]は、同じ命題に関する複数の研究で矛盾する結果が得られている場合に、研究結果に示される統計量を統合し、全体の方向性を導くことができる。EBM では、メタアナリシスは重要な方法論とされており、これを基盤とした系統的総説は強力な Evidence を提供する。しかし、メタアナリシスには問題点もある。効果の認められなかった研究は出版され難いという出版バイアスの問題である。この根本的解決には、未出版の研究結果に関する登録制度が必要である。

一方、データマイニングで得られた複雑な知識構造は、統計量の統合手法であるメタアナリシスで扱うことができない。そこで、分散/並列 KDD の要素技術が期待される。特にメタラーニング[9]の技術は注目される。メタラーニングは学習された知識から学習することと定義される。Prodromidis らが示した概略（Fig.1）は次のようなステージである。（１）トレーニング用データセットからベースレベルの分類システムを学習させ、（２）学習済み分類システムに検証用データセットを与えて予測を行わせ、（３）それらの予測と、検証用データセットからメタラーニング用のトレーニングデータを作成する。(4) このメタラーニング用のトレーニングデータから、メタレベルの最終的な分類システムを得るというものである。学習結果の統合の方法には、さらなる検討の余地があるものの、異なるトレーニング用データで学習し、検証用データを通して知識を統合していくというフレームは、研究者・研究施設を超えて知識や仮説を評価する機会を提供する。すなわち、一次的分析

で得られた知識の後処理システムとしての機能が期待される。個々の施設で実施される医学研究には患者の選択バイアスが存在し、その回避は困難である。多施設で相互検証・知識の統合が可能となれば、選択バイアスの評価・回避がある程度可能となるであろう。

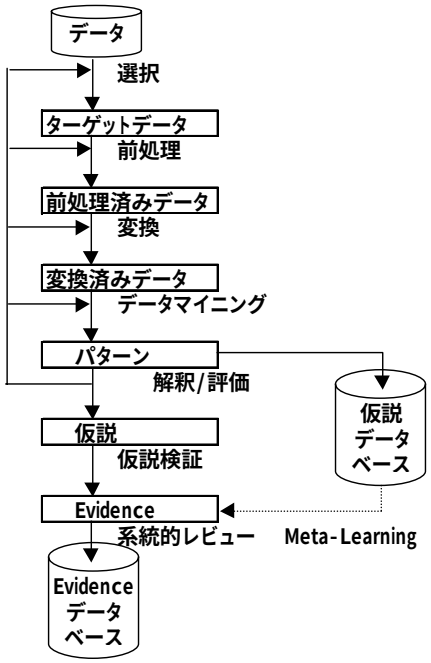


Fig. 2 : EBMにおけるKDDプロセス

前節では、データマイニングによって生成された知識が、Evidence としての利用に耐え得ないこと、観察的研究の結果が正しい確率が 10%にも満たないとの指摘に触れたが、メタラーニングの仕組みが多施設参加の下で実現されれば、仮説の効率的な洗練化が図られ、Evidence として利用可能な信頼性の確保と仮説の的中率の増加がもたらされるものと考えられる。このようなメタラーニングの仕組みは、コクランライブラリーに代表される厳格な Evidence のデータベースに準ずる、Evidence 予備軍としての仮説データベースの役割を担えるであろう。EBM を支える新しい方法論となる可能性が高い。

EBM の実践とエージェント技術　次世代インターネット技術として研究が進められている技術にセマンティックウェブがある[20]。各々のウェブページにコンテンツの意味を表す構造を与え、意味やルールを自由な表現で記述する言語を用いることで、ウェブの知識ベース化を目指している。この技術が実用化されれば、世界中の Evidence を収集するエージェントの設計が可能となる。ユーザーが検索キーを入力するだけで、エージェントが自動的に関連文献を収集し、メタアナリシスを実施し、最良の Evidence を示すことが可能となるであろう。現在、EBM への批判として、関連文献の検索・収集の煩わしさが挙げられている。エージェント技術はこうした障壁に対するブレイクスルーとなる可能性が高い。

EBM における KDD プロセス　本章では３章での議論を整理し、EBM における KDD プロセスを提案する。Fig. 2 にその概要を示す。データの選択からデータマイニングまでのプロセスは Fayyad が示したものと同様である。通常の KDD では、パターンの解釈・評価を経て知識が獲得されることになるが、医療応用においては知識と言うより仮説として認識される必要がある。仮説は仮説検証のプロセスを経て診療に利用可能な Evidence となる。この検証プロセスは、仮説に基づき新たにデザインされた前向き研究か、あるいは実験的研究である。また一方、メタラーニングシステムを前提とした仮説データベースが整備されれば、得られた仮説を登録し、類似したデータマイニングを実施した研究者・研究施設との間で、仮説の相互検証が可能となる。ここで検証された仮説は、前向き研究や実験的研究による検証に比べてその信頼性は低いものの、独立した複数のデータ群から得られた仮説が比較され、バイアスが評価されたという意味で、Evidence としての有効性を議論する価値がある。また、メタラーニングでは前向き研究や実験的研究では扱い難い複雑な仮説も扱うことができるであろう。メタラーニングのプロセスが受け入れられるために、今後多くの議

論が必要である。最後のプロセスでは、個々の Evidence が専門家の系統的レビューによって、より確かな医療情報としてデータベース化され、臨床家に提供される。以下に，a.の論点を整理する。

- データマイニングは一般に観察的研究の立場であり、交絡因子・バイアスの調整がされていないデータを扱うことから、得られる知識は仮説の域を脱することができない。仮説を Evidence まで磨き上げるには、別の検証的研究を追加する必要がある。
- データマイニング技術は、仮説生成ツールとしての利用において有用である。生体データをモデル化するには複雑系としての視点が必要であり、複雑な知識構造の探索による既存の医学知識の再評価にデータマイニング技術は貢献できる。
- 医療応用における倫理的側面において、知識の解釈・評価のプロセスが重要である。このプロセスを支援する技術の開発が望まれる。
- 分散/並列 KDD の要素技術は、一次的分析で得られた知識の後処理システムとして応用が期待される。多施設参加によるメタラーニングの仕組みは、新しい EBM の方法論となる可能性が高い。

EBM は未だ手探りの状態であり、言葉のみが先行しているとの見方もある。EBM の実践を支える情報技術として、データマイニング技術をはじめ人工知能関連技術の積極的な応用が期待される。

b．主観的発見についての検討内容

決定論システムとチャンス発見　ある入力 X に対して，一意に出力 Y を得るシステム f を決定論的システムと呼ぶ．ただし，システムそのものの属性として通常扱われる「状態」も X に含めてのことである。

途中の推論や情報処理の過程が非決定論的，すなわち複数の選択肢を残しながら進めるような場合でも，最終出力がひとつに決まるならばこれを決定論的システムと呼ぶ.

$$Y = f(X). \qquad (1)$$

ここで，式(1)の入力 X は，次元数 n を入力の成分数としてベクトル$(x1, x2, ... xn)$で与えられる．いま，出力 Y を制御する方法を考えよう．このために X の既知成分 m 個を，一般性を失わずに $X0:(x1, x2, ... xm) (m<=n)$と書く．このとき，$f$が既知で $m=n$ の場合が定型的(Programmed)意思決定，それ以外すなわち f が未知または $m<n$ の場合が非定型的(Non-Programmed)意思決定[24]にあたる．決定論的システムをコントロールするために従来行われてきたシステム同定では，後者のうち fが未知で $m=n$ の場合に fを求める問題が主に扱われてきた．比較的単純で影響要因のよく分かっている環境で動作する人工システムには当てはまる定式化である.

ところが昨今，f が未知で m が n より小さいような決定論的システムを考えなくてはならないことが多くなった。例えば，情報を取り入れて行動するまでのヒトは，決定論的システムである．しかし，このシステムは複雑で未知の要因が非常に多く入力として影響するので，fが未知でかつ $m<n$ である．それ故，ヒトというシステムは同定も制御も困難を極める．従来の社会調査で頻繁に行われてきた方法に，選択式あるいは記述式の質問を一定個数与え，これらの質問に対する回答と回答の間の相互関係を計算するパス解析手法が

ある[25]．これは $m=n$，すなわち，ヒトの行動要因は有限個数の質問をすればほぼ全て分かるという近似をヒトの集団について導入して f を同定することにあたる。この近似は，大規模な集団では多様な個性が相殺しあうか，または統計的にはノイズとして扱えるという仮定のもとでは正当性を確保できる．なぜなら，個性を排除して想定される「平均的なヒト」というものは，その行動が全て既知要因で説明できる人為的な架空の決定論的システムであるから。

ところが近年，インターネットを通して自己主張が容易に公開できるなどの情報インフラの変化に伴い，個性が直接社会に影響する可能性が強まり始めた．情報伝送容量の増加に従って，この傾向はさらに広まり，個人間のピアトゥーピアの活発な情報流通が社会を動かすことになろう．個性を無視するのと逆に，社会に影響する個性を，仮に少数であっても理解するような社会調査あるいはマーケティングの手法が，青少年を理解して犯罪でなく社会貢献に導く教育者の意思決定、あるいは新鮮な個性に答える販売者にとって重要である。また，逆に彼ら青少年が少数であっても重要なチャンスを捕らえて将来計画をたてることも大切である。チャンス発見が扱っているのは，この例のように頻度が低くても意思決定において重要となる事象を理解する問題である。

チャンス発見とその基本戦略　チャンス発見は,「意思決定を左右する事象・状況（チャンス）に気付き理解すること」と定義される。たとえば,

- 避難すべき地震の前兆
- 社会的に重要な少数意見，リーダーの発言
- 市場を揺るがす新商品の出現
- 病気の症候

などである。リスクは不の利得を持つが，このチャンスの定義では含めて考える。

まだあまり気付かれていないチャンスの背景には，人の意思決定に影響を与える未知な要因が潜んでいる。そこでは，非定型的な二つの決定論的システムが絡み合っている。ひとつは環境 $f1$ であり，もうひとつは人 $f2$ である。式(2)で，$Y1$ が環境の状態であり，環境の中で $Y1$ が一意に決定された結果として生起し，その一部は人によって観測され，さらにその一部 y は行動への意思決定に利用される。$Y2$ はそうして得られる人の行動である。$x1i, x2i, \ldots xmi$ （$i=1, 2$）が既知な要因であり、$xm(i+1), xm(i+2), \ldots xn$ （$i=1, 2$）は未知な（存在が人の意思決定に反映されていなかった）要因を表す。

$$Y1 = f1(x11, x21, \ldots. xm1, xm1+1, \ldots.xn1). \quad (2)$$

$$Y2 = f2(x12, x22, \ldots. xm2, xm2+1, \ldots.xn2, f1). \quad (3)$$

この中でチャンス発見の問題は、良い y を見出す問題である。すなわち，$Y1$ の一部 y から $f1$ を同定し，$X1:\{x11, x21, \ldots. xm1, xm1+1, \ldots.xn1\}$の値を人が行動 $Y2$ によって操作して別の状態 $Y1$ を生起させる。このとき、実際に操作される変数の集合を $X1c$（$X1$ の部分集合となる）とする。一方、人がこの操作を行う行動 $Y2$ は、その人の関心や考慮している環境状態を含む変数集合 $X2:\{x12, x22, \ldots. xm2, xm2+1, \ldots.xn2\}$の部分集合 $X2c$ に影響され，その人の特性 $f2$ を反映した結果であるが、$X1c$ はこの意思決定者である人が参照し変更を加える変数集合であるから $X2c$ に含まれていることになる。チャンス発見は、Y2 の結果とし

て $Y1'$が当該の人にとって大きな利得を生むことを目的としている。以上を簡単にまとめると，次の問題となる。

[チャンス発見問題 C：C1 を解き，つぎに C2 を解け。]

C1: 事象 y を観測可能な事象の集合 $Y1$ から選び、これを $X1$ と照らして $f1$ を同定せよ。

C2: 行動 $Y2'$により $X1c$ を操作し，その結果 $f1$ により状態 $Y1'$が生起することによる効用 $u(Y1', Y2')$すなわち $\mathrm{gain}(Y1')-\mathrm{cost}(Y2')$を最大化せよ。

$$Y1' = f1(X1c \rightarrow X1c'), \qquad (4)$$

ただし $X1c'$は、それぞれ $X1c$ の操作後の値である。 $f1(X1c \rightarrow X1c')$は変数の部分集合 $X1c$ の値を $X1c'$に変更する操作によって生起する新しい状態 X1 である。

この問題の難しさは、主として次のような悪設定問題である点に由来する。

a. 観測可能な事象集合Πが与えられない。すなわち，計算機に網羅的に与えるには，実世界は広すぎるというフレーム問題である。

b. X1c, X2c が, 未知変数を含む可能性がある。すなわち, $\{x11, x21, ... xm1\}$や$\{x11, x21, ... xm1\}$と一致する保証がない。これもフレーム問題となっている。

c. 関数 $u, f1, f2$ が未知である上, $f2$ は $f1$ を変数とする汎関数であるから不確定さが強い。

現実にチャンス発見手法を創る上では，フレーム問題を厳密に解決するというような究極的に不可能[26]なことは狙わず，近似的に問題 C を解く方が妥当である。しかしながら，チャンス発見の場合，チャンスというものがまだ人に気づかれていない未知要因の影響を大きく受けていると考えられる（後述の「U: 気づかれにくさ」参照)。このように問題の未知成分が難しさの本質となる悪設定問題では，未知成分を無視し良設定とみなすような近似では本質を見逃し，従来から成功例がない。そこで，以下のような戦略がチャンス発見の基本となった。

戦略 a: a については，観測可能な対象を，観測すべきと思われる範囲の中で広げていく。$(xm+1, xm+2, \ldots xn)$は計算機に登録されていない属性であるので、いかなる計算機でも単体でこれら n-m 変数の意味する実体を捉えることはできない。したがって，人が環境から発掘するステップを取り入れる。

戦略 b: b については，未知変数を環境と意思決定主体である人の両方から発掘してゆく。

戦略 c: c については，「u: 主体にとって利得とは何か」「$f1$: 環境中にどのような因果関係が隠れているか」「$f2$: 主体がどのような行動を志向するか」を同定，あるいは関数自体を操作してゆく。特に，依存しあう $f2$ と $f1$ の同定のために，チャンス発見を行うプロセスには主体と環境が，実際に相互作用を行うようにする。

これらの戦略こそ、コンピュータだけでなく，人である意思決定主体をより重要な部品としてチャンス発見システムを構築しなければならない理由である。

チャンス発見プロセスの二重螺旋モデル　チャンス発見において，まず最初に重要となるのが，チャンス発見プロセスを設計することである。コンピュータでいえばフローチャートをまず書くことにあたり，上述のように人を部品として含むチャンス発見システムを動

かすためのプロセスを設計するのである。

ところで，戦略 c に述べたように，チャンスというものの良さすなわちチャンスに基づく行動の効用は，人と環境の相互作用の中で獲得されるものであり，初期から一意に定められるものではない。しかし、全くチャンスの良さを定めないことには，チャンス発見のプロセスを設計する指針を立てることができない。そこで[27]では，チャンスを評価するメタな指標を提案した。それらは

- P: 行動の提案可能性（チャンスを基として行動の提案が可能であること）
- U: 気付かれにくさ（ある事象が，新稀性がなくても，意思決定主体にとっての意味がまだ気づかれていないならこれを満たす）
- G: 成長性（P の提案が採択され実行されるようなチャンスであること）

である．P と G を具体的に把握するために，一人ではなく数人のグループを構成し，グループディスカッション（GI）における人と人の間でのチャンスに基づく提案や採択を通してチャンスを選択してゆくプロセスを実現した。さらにこのプロセスを，意思決定環境についてのデータ（**環境データ**：販売者における購買データなど）に対する視覚的データマイニングによって刺激して促進した[28]。視覚的出力を見て少数参加者から主観的に評価されたチャンスの良し悪しが，次第に他の人にも共有されてゆくコミュニティのダイナミクスに注目したわけである．

このような方法は，*f*1 の同定にデータマイニングを用いる点では Fayyad の知識発見プロセス[29]と共通しているが，図 1 のようにまだ見ていない環境や未知要因を直接観察（戦略 a における観測，および b における環境の未知明変数の値の把握）したり，人の関心を新たなチャンスに向ける（戦略 c における *f*2 の操作）ための心的コンテキストの遷移も焦点となる．ここで，意思決定者がチャンスの意義を解釈するところ（図１下半分）は，次の各状態からなる．

- **新チャンスへの無関心：**ある行動に集中する結果，それと一見関係の薄そうな事象がチャンスであるかどうかについて関心を示さなくなった状態.
- **チャンスへの関心：**前回までの行動あるいはその選択で考慮しなかったチャンスについて，自分の意思決定に対して持つ意義を新たに理解したいと感じる状態.
- **チャンスの意義理解：**あるチャンスの自分の意思決定に対する意義を理解した状態.
- **行動の選択・理解：**チャンスを利用した行動を一意に決定，あるいは共同行動者に提案する．実際に行動する場合も含めてこの状態とみなす．というのは，行動を一意に決定した後は新たな情報（エントロピーすなわち選択の曖昧さを減少させるもの）が使用されたといえないからである．われわれの目的は，チャンスについての情報処理過程を解明し利用することにある．

なお，行動というのは広義に，意思決定者が自分の接することのできる外部環境と相互作用を行うことをさす．たとえば，ある商品を実際に売るというのが通常考えられ商いにおける行動であるが，そのような接し方が戦略上できないならば，その商品をモニターユーザに使用してもらい感想をアンケートで収集するのこともここでは行動と呼ぶ．すなわち，ある行為を行動と呼ぶか提案と呼ぶかは，意思決定者のおかれた状況に依存すると考えるのである．

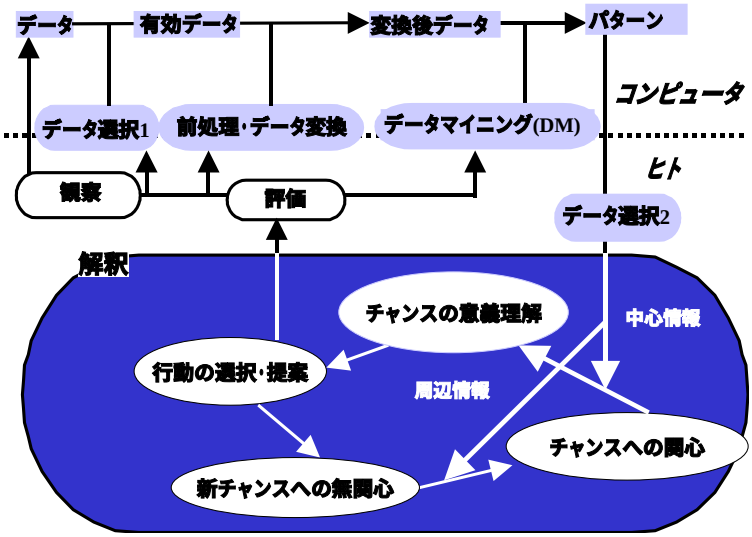


図 1b　チャンス発見プロセスのモデル[27]．[29]では，このうち解釈（Interpretation）は一つのステップに過ぎず，知識の評価（Evaluation）との関係が明らかにされていなかった他，観測（observation）が含まれていなかった．

さて，図１b でもまだ，

- 戦略 b における主体すなわち *f*2 の未知説明変数の発見とその値の観察
- 戦略 c における *u*, *f*2 の同定

は盛り込まれていない。これらは共通して主体に関するデータであるので，主体の思考過程を捕らえたデータ（**主体データ**）をあわせて処理することにより対応したのが図２の二重螺旋モデルである。

この二重螺旋モデルでは，図１b における人の解釈の螺旋過程（関心→理解→行動の決定・提案→関心という繰り返しを経て，チャンスの意義を正しく理解して行くプロセス）と平行して，計算機がデータの受理とデータマイニングを繰り返す螺旋過程を経ながら両者が情報を交換していく．ヒトの思考中も，その思考内容のうち観察できる部分は計算機向けのデータとして蓄積されていく並列動作が、二重螺旋と呼ぶ由来である。

実際，グループインタビュー（GI）における議論の各参加者の考えを，カードに思い当たるだけ書きこんでもらい，そのテキスト内の単語の相関関係を図示する視覚的なテキストマイニング（KeyGraph）の出力図によって議論におけるチャンス表出化を促進するシステムを構築した。このシステムで顧客像を把握した結果，顧客の購買（POS）データからスーパーマーケットの購入金額増加の鍵となる商品や，その店の経営状態のおおまかな変化を示す予兆を発見することができた[31].

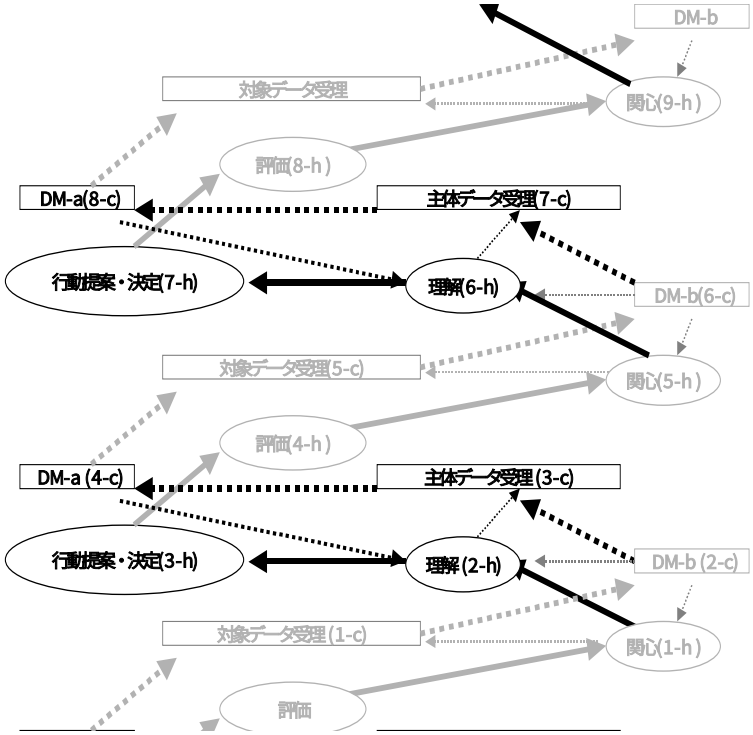


図２　チャンス発見過程の二重螺旋モデル

さらに，社会調査を行った社会学者の「関心」を起点として二重螺旋モデルを実行した結果，社会学的に新規性と説明能力の高い仮説が獲得された[30]。

前者の例について，このモデルに基づき「インターネットが浸透しつつある現代社会において，情報や個性的な価値観が人々の実社会およびネット上の行動にどのように影響するか」の理解を進めることを試みた結果について，次に簡単に示してゆく．

結果（例）：スーパーマーケットの経営分析

某スーパーマーケットの POS データの入手し，これにキーグラフ[32]をかけた。キーグラフの仕組みを簡単に述べると，

1) 生起頻度が高く，かつ近接して出現するアイテム（ここでは，多数の人がセットで答えた回答）の塊を捉え，それぞれの塊を土台と呼ぶ。これらをそれぞれ，黒いノードと、それらの間の黒い色のリンクで表す。
2) 次に，土台中のアイテムと近接して出現することの多いアイテムを屋根と呼び，屋根と土台の間のリンクを柱と呼ぶ。これらは，赤いノードと、それと土台の間の赤いリンクで表される。

図３b ではノードやリンクの色の情報は白黒印刷のため消えているが，以下文中で補う。具体的には，先のキーグラフの仕組みの説明におけるアイテムとして各商品を，近接出現すなわち共起の単位を各顧客の一回の購入商品セットとした。すなわち，同時に買われがちな商品の集合をそれぞれ土台とし，土台中の商品と同時購買される商品を屋根とした。この方法を対象データに対するデータマイニング（図２b の DM-b）として採用し，今回は７名からなるグループワークによって顧客の購買金額の増減の予兆（売り手にとってのチャ

ンス）発見に挑戦した[31]。

最初の「関心」はこの問題への関心であり，POS データ中の商品を全て「ヤサイ」や「ネリモノ」という大分類レベルで表したデータをキーグラフで扱った。一ヶ月あたりの購買金額が上昇している人の POS データだけを扱った結果と，下降している人の結果（図 4 に例示する）を 7 名に見てもらう。次に，7 名それぞれの考えた顧客像を仮説として電子メールで送ってもらい，これらをテキストとして接合したものを主体データとしてキーグラフで処理した。この場合は各単語がアイテムで，各仮説が共起の単位となる。この結果，図 5 を得た。各人はこの図で自分の仮説の該当する位置を図で示し，その位置に応じて 3 組に分かれ，それぞれの考える顧客像を言葉で表現できるようになってから再度で集合して議論した。これは，7 人のうち単に声の大きな人の意見が通るようなことを避け，様々な顧客像をできるだけ考慮して接点を探るために，自分の考えた顧客像を議論者が一旦整理しておく必要があるからである。このような議論を反映しながら環境データすなわち POS データの詳細な取得や理解を進めるプロセスを，二重らせんモデルにしたがって行っていった。

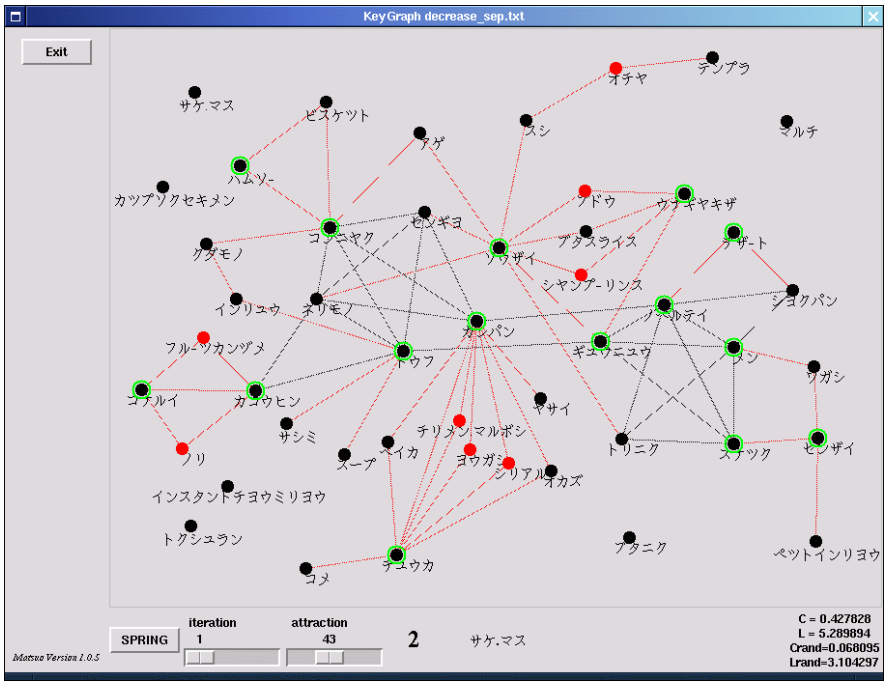


図 4　購入金額の減った人へのキーグラフの結果

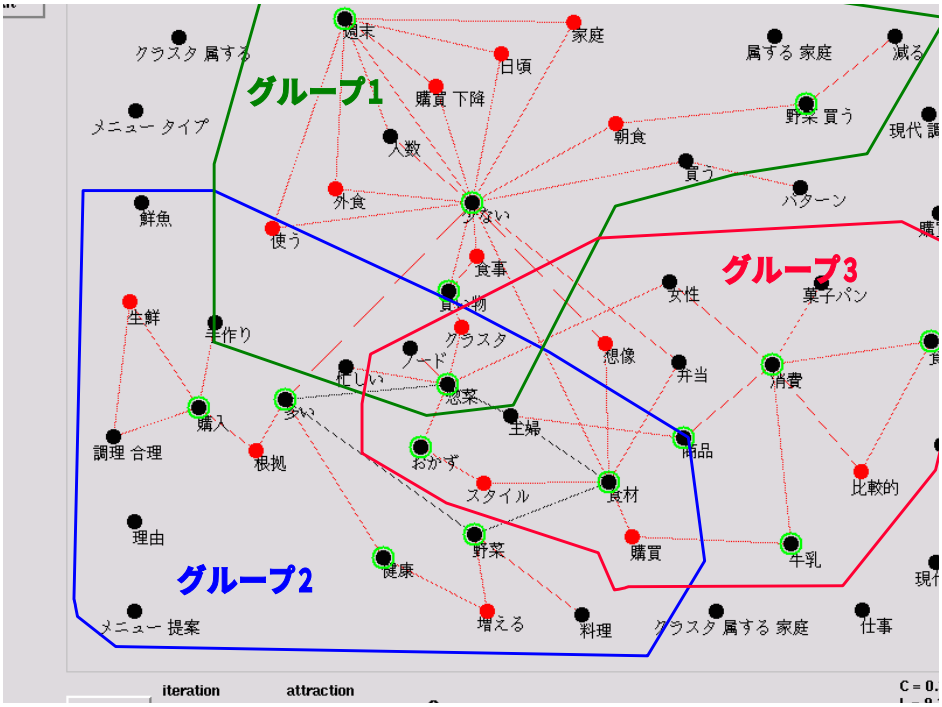


図 5　持ち寄った仮説に対するキーグラフの結果

その結果，次のような意見に一致して到達した。購入金額の増える人は，鮮魚が徐々に食卓の部品として定着して行く人で，地方の特性にこだわる高年齢層が多い。逆に購入金額の減る人は，舶来食（パスタなど）に注意を払いつつもこの店では舶来食を買わなくなる人たちであり，年齢とは別に，嗜好の若々しい層であるという仮説である。

この仮説からすると，スーパーの経営状態は悪化するのではないかと予想された。この予想はおおまかにはその後裏付けられたが，詳細の検証などは事例 1 の新モデル検証とあわせて，今後の課題として稿を改めることにする。ここで伝えたいチャンス発見の新しさというのは，仮説検証の前に行うべき仮説創出の支援を可能にしたことにある。それは未知な要因への気付きを実現するプロセスと言いかえることができる。この「気付き」において，人はシステムの単なるユーザではなく，プロセスの主体である。二重らせんはこのシステムを動かすシナリオであって，キーグラフはその一部品に過ぎない。

考察

データから一般化あるいは発見により抽出された知識を人に提示する場合に、どのような知識が優れているかという指標を、従来は客観的にのみ定義してその指標を最大とするような知識を得る枠組みとなっていた。古典的な手法としては

a. Agrawal らの関連ルール抽出のようにデータの多くを網羅する傾向を一般化するもの

b. Mannila のエピソード解析ように、系列データから頻出パターンを得るもの

など、頻出パターンを選び出すだけのものが代表的である。新奇でもインパクトの大きな事象を発見したりその事象と他の事象との因果関係を理解するための方法として

c. 鈴木（横浜国大）の例外知識発見や Weiss らによる希少事象を予測できる知識の発見

などがある。しかし、これらの知識は、人が予想外の状況に陥った現場において真に知りたいと思う、未知あるいは未認知なる要因を含めた知識にはなっていない。

これは、元々計算機にそのような要因を与えていないため、計算機内部の処理であるデータマイニングでそのような要因を含む知識は生成されないからである。そこで積極的に人やそのグループの主観をシステムに取り入れ、それにより発見される外界の未知要因の重要さを計る指標とその発見に至るプロセスの支援方法を確立するのが本研究の位置付けである。これは、知識発見研究にグループウェアやリスク管理、社会心理学などを統合する学際的方向とも言うことができる。

参考文献

[1] 浜田知久馬: EBMに必要な統計学Ⅰ:EBMにおける統計学の役割, 診断と治療, 88,5, 838/846 (2000)

[2] D.L. Sackett, W.M.C. Rosenberg, J.A.M. Gray, R. B. Haynes, and W.S. Richardson: Evidence-based Medicine: what it is and what it isn't. (editorial), BMJ, 312, 71/72 (1996)

[3] 名郷直樹:EBM実践ワークブック, 南江堂 (1999)

[4] 日本大学医学部公衆衛生教室 homepage, http://www.med.nihon-u.ac.jp/department/public_health/ebm/

[5] The Cochrane Collaboration homepage, http://hiru.mcmaster.ca/cochrane/

[6] National Guideline Clearinghouse homepage, http://www.guideline.gov/index.asp

[7] PubMed homepage, http://www.ncbi.nlm.nih.gov/PubMed/

[8] B. Mullen: Advanced BASIC Meta-Analysis, Lawrence Erlbaum Associates (1989)

[9] A.L. Prodromidis, P.K. Chan, and S.J. Stolfo: Meta-Learning in Distributed Data Mining Systems: Issues and Approaches, Advances in distributed and parallel knowledge discovery, H. Kargupata, and P. Chan (eds.), 81/113, The MIT Press (2001)

[10] 丹後俊郎: 研究の種類に応じたデータのまとめ方:統計的方法適用以前の科学研究者としてのセンス, 日本消化器病学会雑誌, 95, 2, 412/418 (1998)

[11] 神奈川歯科大学 homepage, 医学研究における研究手法: http://www.kdcnet.ac.jp/rinsyo/naika/stat1.htm

[12]津本周作:医学における知識発見手法の可能性,情報処理, 42, 5, 472/477 (2001)

[13] 稲田政則, 金原清子, 五十嵐富三男: 血中酵素データの因果モデルの構築と因子スコアを用いた病態の視覚化, 臨床病理, 47, 1, 61/69 (1999)

[14] 片岡浩巳, 小西修, 西田政明, 杉浦哲郎: 蛋白泳動波形情報のデータマイニングシステム, 日本臨床検査自動化学会会誌, 26, 3, 170/175 (2001)

[15] 岡本康幸, 中野博, 吉川正英, 松岡弘樹, 阪本たけみ, 辻井正: 人工ニューラルネットワークを用いた臨床診断支援システムに関する研究, 臨床病理, 42, 2, 195/199 (1994)

[16] 石野洋子, 寺野隆雄: 模擬育種法と帰納学習を適用したマーケティング情報分析, 人工知能学会誌, 12, 1, 121/131 (1997)

[17] 稲田政則, 寺野隆雄:対話型進化計算と帰納学習による医療データの分析, 第26回知能システムシンポジウム資料, 273/278 (1999)

[18] 松尾収二: 検査の診断能力の評価, 臨床検査情報学,日本臨床病理学会臨床検査情報学専門部会 (編), 34/50, 臨床病理刊行会 (2000)

[19] G.Taubes: Epidemiology Faces Its Limits, Science, 269, 14, 164/169 (1995)

[20] T. Berners-Lee, J. Hendler and O. Lassila :The Semantic Web, Scientific American, http://www.sciam.com/2001/0501issue/0501berners-lee.html (2001)

[24] Simon,H.A., *Administrative Behavior*, (1945) 松田武彦ほか訳『経営行動』ダイヤモ

ンド社 (1965)

[25] Arbuckle, J.L, AMOS (http://www.smallwaters.com)

[26] 松原・橋田：表象なしのロボットもフレーム問題に悩む，現代思想 Vol.18，No.7，pp.160-167

[27] Ohsawa, Y.: Chance Discoveries for Making Decisions in Complex Real World, *New Generation Computing* , Vol.20 No.2 (2002)

[28] Fukuda, H. and Ohsawa, Y., Discovery of Rare Essential Foods by Community Navigation with KeyGraph – An Introduction to Data-based Community Marketing, *in Proc. KES2001* (2001)

[29] Fayyad, U., Shapiro, G.P. and Smyth, P., From Data Mining to Knowledge Discovery in Databases, *AI magazine*, Vol.17, No.3, 37--54 (1996)

[30] 大澤・奈良：チャンス発見プロセスの二重螺旋モデルに基づくアンケート調査データの解析，第 47 回人工知能基礎論研究会資料（2002）

[31]大澤・臼井ほか：OR 学会・マーケティングエンジニアリング部会研究会データ解析コンペティション第一回報告回資料(2001)

[32] Ohsawa, Y. and Yachida, M., Discover Risky Active Faults by Indexing an Earthquake Sequence, in Proc. International Conference on Discovery Science (DS'99) 1999.

参考資料

研究組織

総括班－研究計画(1)

代表者　元田 浩　(大阪大学産業科学研究所)

委員　有川 節夫　(九州大学大学院システム情報科学研究院)

宮野 悟　(東京大学医科学研究所)

山田 誠二　(東京工業大学大学院総合理工学研究科)

北村 泰彦　(大阪市立大学工学研究科)

沼尾 正行　(東京工業大学大学院情報理工学研究科)

山口 高平　(静岡大学情報学部)

鈴木 英之進　(横浜国立大学工学部)

松本 裕治　(奈良先端科学技術大学院大学情報科学研究科)

津本 周作　(島根医科大学医学部)

岡田 孝　(関西学院大学情報メディア教育センター)

大澤 幸生　(筑波大学社会工学系)

有村 博紀　(九州大学大学院システム情報科学研究院)

鷲尾 隆　(大阪大学産業科学研究所)

A01 班：アクティブ情報収集

班代表　沼尾 正行　(東京工業大学大学院情報理工学研究科)

研究項目 A01－研究計画(2)「WWW におけるメタ情報源の獲得」

研究代表者　山田 誠二　(東京工業大学大学院総合理工学研究科)

研究分担者　高間 康史　(東京工業大学大学院総合理工学研究科)

研究分担者　小野田 崇　((財)電力中央研究所情報研究所)

研究項目A01－研究計画(3)「分散動的情報源からのアクティブ情報収集」

研究代表者　北村 泰彦　(大阪市立大学大学院工学研究科)

研究分担者　平山 勝敏　(神戸商船大学商船学部)

研究項目A01－研究計画(4)「多段階学習方式によるデータ収集と前処理の自動化」

研究代表者　沼尾 正行　(東京工業大学大学院情報理工学研究科)

研究分担者　櫻井成一朗　(東京工業大学大学院情報理工学研究科)

研究協力者　市瀬 龍太郎　(国立情報学研究所知能システム研究系)

A02班：ユーザ指向アクティブデータマイニング

班代表　山口 高平　(静岡大学情報学部情報科学科)

研究項目A02－研究計画(5)「構造データからのアクティブマイニング」

研究代表者　元田 浩　(大阪大学産業科学研究所)

研究分担者　鷲尾 隆　(大阪大学産業科学研究所)
研究分担者　矢田勝俊　(関西大学商学部)
研究分担者　Tu Bao Ho　(北陸先端科学技術大学院大学知識科学研究科)
研究分担者　吉田 哲也　(大阪大学産業科学研究所)

研究項目A02－研究計画(6)「メタ学習機構に基づくアクティブマイニング」

研究代表者　山口 高平　(静岡大学情報学部情報科学科)
研究分担者　橘 惠昭　(愛媛大学法文学部総合政策学科)
研究分担者　和泉 憲明　(静岡大学情報学部情報科学科)
研究分担者　大崎 美穂　(静岡大学情報学部情報科学科)

研究項目A02－研究計画(7)「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」

研究代表者　鈴木 英之進　(横浜国立大学大学院工学研究院)
研究分担者　鍾 寧　(前橋工科大学工学部情報工学科)

研究項目A02－研究計画(8)「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」

研究代表者　松本 裕治　(奈良先端科学技術大学院大学情報科学研究科)
研究分担者　新保 仁　(奈良先端科学技術大学院大学情報科学研究科)

A03班：アクティブユーザリアクション

班代表　津本 周作　(島根医科大学医学部医療情報学)

研究項目A03－研究計画(9)「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」

研究代表者　津本 周作　(島根医科大学医学部医療情報学)
研究分担者　平野 章二　(島根医科大学医学部医療情報学)
研究分担者　高林 克日己　(千葉大学医学部附属病院医療情報部)

研究項目A03－研究計画(10)「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」

研究代表者　岡田 孝　(関西学院大学情報メディア教育センター)
研究分担者　比嘉 真弓　(関西学院大学情報メディア教育センター)
研究分担者　高橋 由雅　(豊橋技術科学大学工学部)
研究分担者　加藤 博明　(豊橋技術科学大学工学部)

研究項目A03－研究計画(11)「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」

研究代表者　大澤 幸生　(筑波大学社会工学系)
研究分担者　寺野 隆雄　(筑波大学社会工学系)

活動報告

計画研究代表者会議

第１回計画研究代表者会議

日時：平成１３年7月２１日（土）１３：３０～１６：３０

場所：東京工業大学百年記念館第一会議室

（東京都目黒区太岡山 2-12-1　東京工業大学）

議題：

１．経緯

２．今後の予定

３．ヒアリング報告

４．推進体制について

５．研究計画について

６．総括班予算について

７．研究会の予定について

８．年度末報告会について

９．来年度の国際ワークショップについて

１０．特集号について

１１．その他

第 2 回計画研究代表者会議

日時：平成１３年9月4日（火）１７：００～１８：００

場所：Freiburg 大学，ドイツ

（12th European Conference on Machine Learning:ECML2001 国際会議会場）

議題：

１．今後の投稿及び研究会・ワークショップ開催について

２．プロジェクト推進基本方針について

３．その他

第 3 回計画研究代表者会議

日時：平成１３年9月6日（木）１７：００～１８：００

場所：Freiburg 大学，ドイツ

（12th European Conference on Machine Learning:ECML2001 国際会議会場）

議題：

１．人工知能学会特集号の提案が沼尾先生からあった。

２．プロジェクトホームページ・名簿について
３．プロジェクトが対象とするデータについて
４．その他

総括班会議

第１回総括班会議
日時：平成１３年１１月１２日（月）１０：００～１３：００
場所：はこだて未来大学「５階会議室」　（〒041-8655 北海道函館市亀田中野町 116-2）
議題：
１．今年度の全体計画
２．共通医療データと各班の取り組み
３．各計画研究の今年度の計画と協力体制
４．来年度の計画(国際ワークショップ，研究会など)
５．その他

第 2 回総括班会議
日時：平成１４年3月6日（水）１２：００～１３：００
場所：
議題：
１．本年度の活動の総括
２．来年度の活動計画
　１）共通医療データの取り組み
　２）研究会計画
　３）国際ワークショップ計画
　４）特集，出版計画
　５）公開シンンポジウム
３．その他

班会議

A01 班　第１回　班会議
日時：平成１３年１２月２６日(水)　１３：００～１７：００
場所：東京工業大学大岡山キャンパス:（西 8 号館 E 棟 5 階コラボレーションルーム）:
議題：
１．**沼尾正行**：A01 班の概要について
２．**チャン・ナム・トアン，沼尾正行**：
　EDLINE の解説および医療ネットワークの紹介および議論

３．**沼尾正行，チャン・ナム・トアン，山田誠二，高間康史，小野田崇，北村泰彦，平山勝敏**：各課題グループからの報告
４．**元田浩，沼尾正行**：議論のまとめと今後の方針

A02 班＋A03 班　第１回　班会議

日時：平成１４年1月３０日（水）１３：００～１８：００
　　平成１４年1月３１日（木）　９：００～１２：００
場所：３０日（水）千葉大学医学部附属病院 3 階第二会議室
　　３１日（木）千葉大学医学部附属病院 3 階第一会議室
議題：
１．**里村洋一**：千葉大学医学部附属病院情報部についての紹介
２．各班から肝炎データに対する解析結果発表と**高林克日己，横井英人**からのコメント
３．**横井英人**：第一内科から提供される新しいデータについて
４．第一内科から提供される新しいデータの扱い方について
５．今後の共通医療データ解析に対する班会議について
６．　3 月の公開シンポジウムへの報告書原稿について
７．韓国での情報処理学会研究会について
８．全国大会の AI レクチャーにおけるアクティブマイニング企画について
９．人工知能学会アクティブマイニング特集号について

報告会兼公開シンポジウム

平成１３年度成果報告公開シンポジウム
日時：平成１４年3月5日（火）１３：００～１８：００
　　平成１４年3月6日（水）　９：３０～１６：３０
場所：筑波大学ビジネス科学研究科　大講義室 G501 号
「総括班活動報告」元田　浩（大阪大学産業科学研究所）
「アクティブマイニングと EBM」津本周作（島根医科大学学医学部医学科医療情報学）
「対話的文書検索によるアクティブ情報収集」
　山田 誠二（東京工業大学大学院総合理工学研究科），　岡部 正幸（科学技術振興事業団）
「WWW 上の情報収集/可視化のための免疫ネットワーク・メタファを用いた
　クラスタリング」高間 康史，廣田 薫（東京工業大学大学院総合理工学研究科）
「WWW 情報管理のための Web ページにおける部分情報の更新モニタリング」
　中井有紀，山田誠二（東京工業大学大学院総合理工学研究科）
「静的・動的知識に基づくアクティブ情報収集」
　北村泰彦（大阪市立大学大学院工学研究科）
「伝言ゲーム型の情報収集とデータ前処理」

沼尾正行（東京工業大学大学院情報理工学研究科）
「HTML のリンク構造と構文的特徴に基づく知識獲得について」
桜井成一朗（東京工業大学大学院情報理工学研究科）
「高次元インデックス技術を用いた検索処理性能向上について」
河野浩之（京都大学大学院情報学研究科）
「構造データからの部分構造抽出の高速化」
松田喬，元田浩，鷲尾隆，吉田哲也（大阪大学産業科学研究所）
「ブランド間関連購買の知識表現と評価基準」矢田勝俊（関西大学商学部）
「Document clustering by a tolerance rough set model」
Tu Bao Ho, Saori Kawasaki (JAIST: Japan Advanced Institute of Science and Technology)and Ngoc Binh Nguyen (Hanoi University of Tehcnology, Vietnam)
「慢性肝炎データセットのクレンジングとマイニングの試み」
畑澤寛光，佐藤芳紀，山口高平（静岡大学情報学部）
「スパイラル的例外性発見に向けて」
鈴木英之進（横浜国立大学大学院工学研究院），　鍾 寧 (Ning ZHONG)（前橋工科大学）
「医学生物学論文概要における専門用語の抽出と意味推定」
新保仁，山田寛康，松本裕治（奈良先端科学技術大学院大学）
「ラフ集合に基づく診療情報のアクティブマイニング」
平野章二，津本周作（島根医科大学学医学部医学科医療情報学）
「カスケードモデルの発展と発ガン性・変異原性を示す分子の発見」
岡田孝（関西学院大学）
「分子の構造類似性にもとづくデータマイニング」
高橋由雅、加藤博明（豊橋技術科学大学）
「医療データにおける知識発見とそのヒューマンインタラクション」大澤幸生（筑波大学）

関連学会及び研究会

第 46 回基礎論研究会(SIGFAI) 第 54 回知識ベースシステム研究会(SIG-KBS)合同研究会
テーマ：「アクティブマイニング」および 一般 (http://www.ymd.dis.titech.ac.jp/sig-kbs/)
日時：平成１３年１１月１２日（月）１３：００～１８：３０
平成１３年１１月１３日（火）　８：３０～１８：２０
平成１３年１１月１４日（水）　８：３０～１４：１５
場所：はこだて未来大学

セミナー

ＥＢＭセミナー
場所：大阪大学産業科学研究所本館元田研究室（〒567-0047 大阪府茨木市美穂ヶ丘 8-1）

日時：平成１３年9月２８日（金）１４：３０～１７：００
講師：島根医科大学　教授　津本周作