特 集

分散動的情報源からのアクティブ情報収集

肝炎データマイニングにおける発見ルールフィルタリングの試み

Active Information Gathering from Distributed Dynamic Information Sources

An Approach Toward Discovered Rule Filtering in Hepatitis Data Mining

北村 泰彦
Yasuhiko Kitamura
関西学院大学
Kwansei Gakuin University
ykitamura@ksc.kwansei.ac.jp, http://ist.ksc.kwansei.ac.jp/~kitamura/

keywords: data mining, information retrieval, discovered rule filtering, MEDLINE

1. はじめに

アクティブマイニングは情報収集，データマイニング，ユーザリアクションの技術を融合することにより，利用者の目的にあった質の高い知識の効率的な発見を目指すデータマイニングの新しいアプローチである [Motoda 02]．本稿では情報収集技術の一つとして，インターネット上からの文献情報検索結果に基づき，データマイニング結果をフィルタリングする発見ルールフィルタリング手法 [Kitamura 02, Kitamura 03, Kitamura 04] について述べる．これはデータマイニングの対象となる情報源と，動的に変化するインターネット上の情報源を高い次元で統合することで，質の高い知識の発見を支援する試みである．

データマイニングは大量のデータを機械処理することにより，利用者にとって有用な知識を自動的に発見しようとする手法である．一般的には与えられたデータに含まれる属性間の関係から統計的に意味のある関係を発見しようする．しかしながら単に統計的な特徴だけでデータマイニングを行うなら，利用者にとって興味のない，既知の知識が大量に得られる可能性がある．発見された知識が既知であるかどうかの判定は，当然のことながら，マイニングの対象となっているデータの解析から得られることはなく，外部の情報源が必要になる．そこでわれわれは外部の情報源としてインターネットを介して利用可能な文献データベースを用い，その検索結果に基づき発見された知識（ルール）のフィルタリングを行う発見ルールフィルタリングを研究している．

発見ルールフィルタリングにはミクロビューとマクロビューの二つのアプローチが考えられる．ミクロビューアプローチは発見ルールに直接関連のある文献を検索し，その文献数に応じてフィルタリングを行う．このアプローチでは利用者は，フィルタリングされた発見ルールだけでなく，発見ルールに関連する文献を同時に入手することができ，それが新たな知識発見へと結びつく可能性がある [Kitamura 02]．一方でその成功のためには，精度の高い情報検索技術が必要になり，課題も残されている．これに対してマクロビューアプローチでは発見ルールに関連する研究活動の傾向を大まかに観察し，その結果に基づきルールのフィルタリングを行うものである．このアプローチは情報検索の精度がある程度低くても利用可能であり，また発見ルールに関連する文献数が多い場合にも有効である．

本稿では以下，発見ルールフィルタリングの原理と手順について述べた後，その具体的手法としてミクロビューアプローチとマクロビューアプローチについて述べ，肝炎データマイニングの領域におけるその有効性について述べる．

2. データマイニングと情報検索の統合

データマイニングとは，複数の属性集合 $A_1, A_2, \cdots, A_n$ に対し，それらの関係を示す大量のデータ集合 $D(\subseteq A_1 \times A_2 \times \cdots \times A_n)$ から特徴的な属性間の関係を発見することと定義できる．（ここでは簡単のために各属性値は 0 あるいは 1 の値を取ると仮定する．）すなわちデータマイニングはデータ集合 D を入力とし，属性間の関係を表すルール集合を出力とする関数 $dm(D) \subseteq R(= \{< A_{c1}, A_{c2}, \cdots, A_{cm} \rightarrow A_d >\})$ として定式化できる．このようなルール集合を求める手法としては一般的には正答率 (precision) や再現率 (recall) を考慮する統計的手法が用いられることが多い．ただし，新奇なルールを発見しようとするシステムでは再現性を犠牲にした手法がとられることもある．

一方，情報検索とは多数のキーワード集合 $B_1, B_2, \cdots, B_m$ が与えられているときに，それらを含む大量の文献集合 $D' \subseteq B_1 \times B_2 \times \cdots \times Bm$ からキーワードの共起数を求めることとして定義できる．すなわち情報検索は文献集

合とキーワード集合を入力とし，キーワードの共起数を出力とする関数 $ir(D', \{B_{k1}, B_{k2}, \cdots, B_{kp}\}) \in \mathbf{Int}$ として定式化できる．（ここで **Int** は整数の集合である．）実用上，情報検索では共起数そのものよりも，キーワードを含む文献リストが出力となる．

それではデータマイニングと情報検索を組み合わせることによりどのようなことが可能であろうか．まずデータマイニングにおける属性 A_i を情報検索におけるキーワード B_j に関連付ける関数 $c(A_i) = B_j$ を得ることができるなら，データマイニング結果と情報検索結果を関連付けることが可能になる．例えば，データマイニングの結果としてルール $< A_{c1}, A_{c2}, \cdots, A_{cm} \rightarrow A_d >$ が得られたとしよう．このルールを構成する属性に関連するキーワードを用いて情報検索を行うと共起数 k が得られる．すなわち，$ir(D', \{c(A_{c1}), c(A_{c2}), \cdots, c(A_{cm})\}) = k$ である．このとき k の値の大きさに応じて発見ルールのランク付けを行うことができる．k が非常に大きな数値であれば，発見されたルールは既知のものである可能性が大きいし，その逆であれば未知の可能性が大きい．

情報検索にはさらに付加的なキーワードやパラメータを追加することも可能である．例えば，ある文献情報検索システムでは文献が出版された年を入力とした検索が可能になっている．これより発見されたルールが過去のトピックであるのか，最新のトピックであるのかを識別することが可能になる．また，利用者が興味を持つ領域を表すキーワードを付加すれば，それを含めた評価も可能になる．

3. 発見ルールフィルタリングの手順

発見ルールフィルタリングではデータマイニングシステムにより発見されたルールに対して，それに関連する情報をインターネット上から検索し，その結果に基づき，発見ルールのフィルタリングを行う．ここでの議論をより具体的なものとするために，肝炎診療データからのマイニングを知識発見の対象領域として議論を進める．発見ルールフィルタリングの具体的な手順は以下の通りである．

3·1 発見ルールの獲得

第 1 段階はデータマイニングシステムを利用してルール形式の知識を獲得することである．静岡大学のグループでは，千葉大学医学部より提供された肝炎データから，肝炎の進行具合を示す血液データ (GPT) と他の検査データとの時系列的な相関関係に着目し，様々なルール形式の知識発見が行われている [Ohsaki 02]．その一例は図 1 に示すようなものである．

これは GPT（グルタミン酸ピルビン酸トランスアミナーゼ）と TTT（チモール混濁試験），D-BIL（直接ビリルビン）の関係を表すものであり，「24ヶ月間，D-BIL が一定で，TTT が減少し，GPT が増加しているなら，その後 24ヶ月間の GPT は減少する」というルール表記を行うことが可能である．

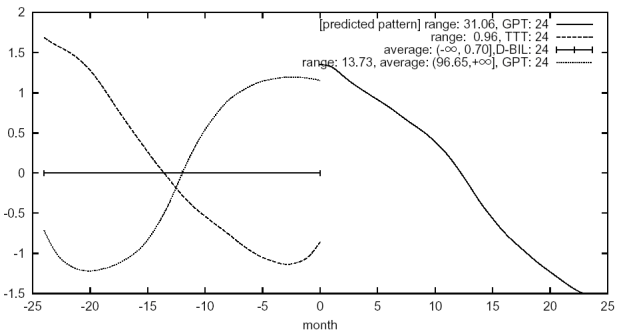

図 1 発見ルールの一例

3·2 発見ルール駆動情報検索

第 2 段階では，発見されたルールに関連する情報をインターネットを利用して収集する．肝炎に関連するインターネット情報源として一般の Web 情報源はあまりにも雑多な情報が混在しているので，本研究では医学・生物学関係の文献データベースである MEDLINE を用いている．MEDLINE（MEDlars on LINE）は，米国をはじめ 70 カ国で出版された 4000 誌を超える医学・生物学系学術雑誌からのアブストラクトを含む書誌情報データベースであり，1966 年以降の 1100 万件以上のデータが蓄積されている．PubMed[*1] は NCBI (National Center for Biotechnology Information) によりインターネット上に無料提供されている MEDLINE の検索サービスであり，一般の検索エンジン同様，キーワードを入力することにより，MEDLINE 文献の検索を行うことができる．PubMed ではさらに出版年別の検索を行うことも可能である．また近年は PubMed の Web サービス ESearch[*2] も運用が開始されている．検索結果は XML 形式で出力することができ，その処理が容易である．

発見ルールに関連する情報を MEDLINE データベースから検索するためには，それに関連するキーワードを獲得する必要がある．このキーワードは発見ルール，マイニング領域，利用者の興味から抽出することが可能で，以下のように分類される．

(1) 発見ルールを構成する属性に関連するキーワード：発見ルールから直接抽出されるキーワードであり，発見ルールを構成する属性名がそれに該当する．例えば図 1 に示される発見ルールから得られるキーワードとしては GPT, TTT, D-BIL が該当する．ただし発見ルールに含まれる属性名には略称が用いられていることが多く，一般名への変換も必要になる場

*1 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
*2 http://eutils.ncbi.nlm.nih.gov/entrez/query/static/esearch_help.html

合もある．例えば，TTT は thymol turbidity test, GPT は ALT に変換する．

(2) 発見ルールを構成する属性の性質に関連するキーワード：発見ルールがある属性値の増加や減少について述べている場合があるが，これらを表すキーワードとして increase や decrease を含めることも可能である．ただ発見ルールが複数の属性から構成される場合には，そのキーワードがどの属性に関するものであるのかを区別することは難しい．例えば，図 1 に示される発見ルールにおいて，キーワードだけを用いて TTT の減少と GPT の増加を示す文献のみを検索することは難しい．そのために文献検索には属性キーワードだけを用いて，その性質の記述に関しては自然言語処理手法を用いて確認することが望ましいと考えられる．

(3) 領域に関連するキーワード：データマイニングの目的や背景を表すキーワードである．これは固定的なキーワードとしてあらかじめ用意しておく．例えば，肝炎データマイニングの場合には hepatitis（肝炎）といったキーワードがこれに該当する．

(4) ユーザの興味に関連するキーワード：発見知識に対するユーザの興味を表すキーワードである．このようなキーワードを獲得する方法としては，直接的手法としてユーザから直接獲得する方法と，間接的手法として 3･5 節に示すように，適合性フィードバック手法 [Onoda 02] により間接的に獲得する方法が考えられる．

3･3 知的情報収集

発見ルールより抽出されたキーワードの組み合わせを用いて PubMed より MEDLINE 文献検索を行う．このような文献検索を発見ルールの数だけ繰り返す．当然のことながら発見ルールの数が増えるにつれ PubMed への検索の回数も増加することになる．一方，PubMed は世界中の多くの研究者に公開されている検索システムであり，その負荷をできるだけ少なくするような工夫が必要である．現実に，あまりに多くの検索を連続的に行うと検索サービスの利用が打ち切られてしまうこともある．そこで，図 2 に示すように，過去の検索の履歴を保持し，意味のない文献検索や冗長な検索を行わず，効率的な情報検索を行うことが望ましい．この図では検索キーワードとその検索文献数が示されている．例えば，この履歴を参照すれば hepatitis, gpt, t-cho をキーワードとして含む文献検索は結果が 0 になることがわかる．

なお，MEDLINE は日々新たな文献が登録されので，同じキーワードであっても定期的に再検索する必要がある．このような動的な情報源から効率的に情報収集する手法としては [北村 01] があげられる．

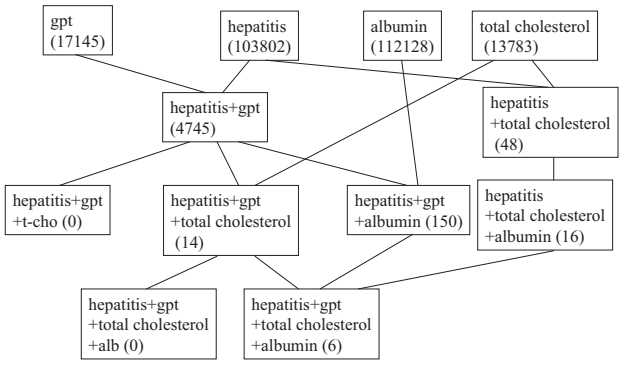

図 2 検索履歴の保存

3･4 発見ルールフィルタリング

MEDLINE 文献検索の結果に応じて発見ルールのフィルタリングを行う．具体的には文献検索結果に応じてルールのランキングを行ったり，ある閾値を設定して，それ以下のものを排除したりする．文献検索結果とランキングをどのように対応付けるかは本研究の中核となる重要な研究課題であるが，以下のような仮説を前提とすることができるであろう．

- 文献数が多ければ，それだけ既知のルールといえる．
- 出版時期の新しい文献が多ければ，それだけ既知のルールはホットな話題を扱っているといえる．

発見ルールフィルタリング手法に関しては 4 章でより詳しく述べる．

3･5 適合性フィードバック

発見ルールに関連するキーワードを元に文献検索を行うだけでは，かなり広い範囲の文献にヒットする可能性があり，その中には発見ルールと関連していても，ユーザの興味とあまり関連していないものが含まれることもある．この問題に対処する方法としては，ユーザがその興味を示すキーワードを直接入力することが考えられるが，これはユーザにとって負担になる場合もある．そこで間接的にユーザの興味に関連する文献を獲得する手法として適合性フィードバック [Onoda 02] がある．適合性フィードバックは検索された文献に対してユーザが自らの興味と関連があるかどうかを Yes/No でフィードバックする手法である．システムはユーザからのフィードバックを手がかりに，文献アブストラクトを解析し，興味ある文献とそうでない文献に分類できる．この結果を用いることにより，発見ルールフィルタリングの結果にユーザの興味を反映させることができる．

4. MEDLINE 文献検索に基づく発見ルールフィルタリング

MEDLINE 文献検索の結果を用いて発見されたルールをいかにフィルタリングするかは本研究の最も重要な

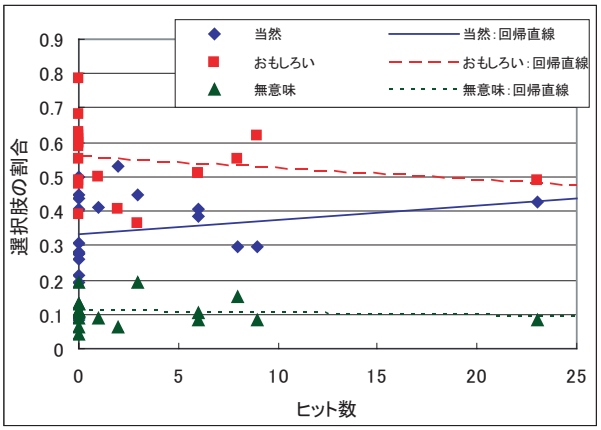

図 4 ミクロビューアプローチの評価

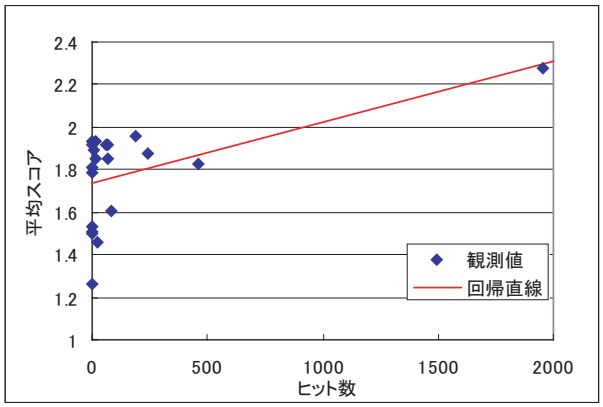

図 5 キーワード数 2 の場合におけるヒット数と評価の関係

課題である．これにはミクロビューアプローチとマクロビューアプローチの二つが考えられる．

4·1 ミクロビューアプローチ

ミクロビューアプローチは発見ルールに関する文献検索のヒット数に応じて，発見ルールのランク付けをしようとするアプローチである．ミクロビューアプローチでは以下のような仮定をおいている．

(1) 既知のルールのヒット数は多い．
(2) 未知のルールのヒット数は少ない．
(3) ゴミのルールのヒット数はゼロである．

これは，既知のルールはよく研究がなされており，それに関連する文献は多く出版されているという仮定に立っている．ヒット数が少なければ，それは十分な研究がなされているとは限らないので，未知ルールに分類することができると仮定する．一方で，ルールが無意味なものに関しては誰も研究しないのでゼロであると仮定している．この仮定では，既知か未知かの境界は曖昧であるが，実際には何らかの閾値を用いて区別することになる．

以上の仮説の妥当性を検討するために，医学生へのアンケートに基づく予備評価を行った． アンケートの質問文を図 3 に示す．問 1 の質問項目には，静岡大学のグループにより得られた 30 個のルールの中からランダムに 20 個選択し，そこに含まれる属性名を使用した．アンケートの対象は，医師国家試験を直前に控えた大阪市立大学医学部の 6 回生，47 人である．国家試験を直前にひかえた医学生であることから，医学的な常識とも言える教科的な知識についてはかなり詳しいと考えられる．

図 4 に問 1 のアンケート結果と文献検索結果をまとめたものを示す．それぞれの発見ルールに対して，三つの選択肢が選ばれた割合と関連文献検索したときのヒット数の関係をプロットしたものである．またそれぞれの選択肢について求めた回帰直線も表示している．それぞれの選択肢について，その回帰直線の傾きを有意水準 95 %で t 検定したが，アンケート結果とヒット数の間には有意な相関を見出すことはできなかった．

ミクロビューアプローチによる検索結果と医学生の評価との間に相関が得られなかった理由としては，単純なキーワード検索では検索結果にノイズが多く含まれ，検索ヒット数は発見ルールとの関連よりも，検索するキーワード数に依存することにある．よって情報検索の精度が悪い場合にはミクロビューアプローチは十分に機能しないことが明らかになった．

ただし，キーワード数が少ない場合には医学生の評価とヒット数に相関があることが示された．図 3 の問 2 に示されるアンケート調査を実施した．問 2 は，発見ルールに含まれるキーワードの中から 20 組のキーワード対を用いて作ったものである．図 5 に問 2 の結果を示す．「強い関係がある」,「関係がある」,「あまり関係はない」,「無関係」のそれぞれに 3，2，1，0 のスコアをつけ，その平均と文献検索によるヒット数をプロットした．平均スコアとヒット数との間の相関係数は 0.54 であり，ヒット数と医学生の評価には相関があることがわかった．したがって文献検索に用いるキーワード数を 2 に限定すれば，検索結果と医学生の評価には相関があり，この結果は次節で述べるマクロビューアプローチへと展開されてゆく．

先にも述べたように，精度の悪いキーワード検索に基づくミクロビューアプローチは発見ルールフィルタリングのためには十分有効であるとはいえないが，ミクロビューアプローチでは発見ルールに関連する文献を検索し，それを利用者に提示することができる．これは利用者に対して新しい知見を生み出す助けとなる．以下に予備実験の過程で得られたその一例を示す．

まず，利用者（医者）に図 1 に示す発見ルールを提示し，コメントしてもらった（コメント 1）．次に，上記の手法を用いて関連文献を検索した．キーワード GPT と TTT を用いたところ検索結果は 11 件であった．その中には図 6 に示されるような利用者が興味を持つ文献があり，それに関してコメントしてもらった（コメント 2）．

コメント 1 は以下のようなものである．

問１：下のキーワードの組み合わせで MEDLINE を検索するとしたときに，検索結果がどのようになると思いますか？次の三つの中から選んでください．

1. 当然（どのような内容の文献が見つかるか，ある程度予想できる）
2. おもしろい（自分が知らないことが載っている文献が見つかるかもしれない）
3. 無意味（ほとんどヒットしないだろう）

（１） ALT TTT （1. 当然、２．おもしろい、３．無意味）
（２） TTT 直接ビリルビン ALT （1. 当然、２．おもしろい、３．無意味）
（３） ALT 総コレステロール C 型肝炎 （1. 当然、２．おもしろい、３．無意味）
…

問２：次の検査項目同士の関係について，次の四つの中から適切だと思うものを選んでください．

1. 強い関係がある．
2. 関係がある．
3. あまり関係は無い．
4. 関係ない．

（１） ALT 総ビリルビン （ 1、 2、 3、 4 ）
（２） ALT 総コレステロール （ 1、 2、 3、 4 ）
（３） ALB 総コレステロール （ 1、 2、 3、 4 ）
…

図 3 アンケート

1: Hepatol Res 2001 Sep;21(1):67-75

Comparison of clinical laboratory liver tests between asymptomatic HBV and HCV carriers with persistently normal aminotransferase serum levels.

Murawaki Y, Ikuta Y, Koda M, Kawasaki H.

Second Department of Internal Medicine, Tottori University School of Medicine, 683-8504, Yonago, Japan

We examined the clinicopathological state in asymptomatic hepatitis C virus (HCV) carriers with persistently normal aminotransferase serum levels in comparison with asymptomatic hepatitis B virus (HBV) carriers. The findings showed that the thymol turbidity test (TTT) values and zinc sulfate turbidity test (ZTT) values were significantly higher in asymptomatic HCV carriers than in asymptomatic HBV carriers, whose values were within the normal limits. Multivariate analysis showed that the independent predictor of serum TTT and ZTT levels was the HCV infection. In clinical state, simple and cheap tests such as TTT and ZTT are useful for mass screening to detect HCV carriers in medical check-ups of healthy workers.

PMID: 11470629 [PubMed - as supplied by publisher]

図 6 発見知識に関連する文献

「TTT は抗体の働きを示す指標であり，抗体の働きが活発になればそれに応じて肝炎も緩和されるので GPT は減少する．したがってこのルールは常識的な知見である.」

すなわち発見されたルールは医学的には常識の範囲内の知見であるといえる．次に，コメント 2 は以下のようなものである．

「これは肝炎ウイルスのキャリア（発症していない人）において（安価な）TTT 検査で B 型ウイルスと C 型ウイルスの比較ができるという内容であり，TTT と肝炎との関連を述べた新しい文献（2001 年）である．ただし報告例は小規模な事例数しか扱っていない．発見ルールはこれがキャリアではなく，発症した人に見られることを指摘しており，発見ルールはこの論文を臨床データの観点から支える意味で重要である.」

この予備実験により明らかになった効果は，利用者に対して発見ルールに関連する最新の文献を提供しただけでなく，発見ルールに対する異なる視点を与え，新たなマイニングプロセスのきっかけを与えているという点である．すなわち発見ルール単独の提示では，常識的な知見としか認識されなかったものが，それに関連する文献を同時に示すことにより，新たな研究の動機（肝炎の種別に応じた TTT の違いを，より大規模な肝炎データを用いて実証する）を生じさせる原因を作っている．これは新しいマイニングプロセス（肝炎の種別と TTT の関係に絞り込んだデータマイニング）のきっかけを与えることに成功しているともみなすことができる．

4·2 マクロビューアプローチ

ミクロビューアプローチはキーワード検索を用いて発見ルールに直接関連する文献を検索し，そのヒット数に応じてフィルタリングを行うアプローチであったが，精度の悪い検索では十分なフィルタリング能力を発揮することはできなかった．ヒット数は発見ルールの新規性よりも検索キーワードの数に依存することが明らかになった．一方，検索キーワードを 2 に限定した場合には，新規性との相関が見られた．これに基づき，マクロビューアプローチでは発見ルールから抽出されるキーワードの全ての対に関する共起文献数を参照することで，発見ルールに関連する研究動向をおおまかに推測しようとするものである．

図 7 と図 8 は 2 つのルールに関する共起文献数をグラフとして表したものである．グラフ中でのノードはそれぞれのキーワードを表し，エッジの長さは二つのキーワードでの共起数の逆数，添字は共起数を表す．

図 7 のルールでは，ALB，GPT，T-CHO のいずれの対に関しても共起数が多く，それぞれ関係が強い属性であることがわかる．図 8 のルールでは，GPT, female, G-GTP 間には相互に関係が強いが，hemolysis と G-GTP 間と blood group a とその他のキーワードとの間には関係が弱い．

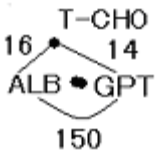

図 7 GPT, ALB, T-CHO 間の共起関係

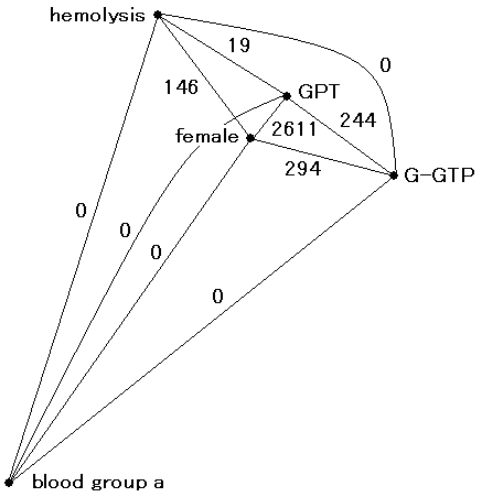

図 8 GPT, G-GTP, hemolysis, blood group a, female 間の共起関係

それぞれのルールを明確に特徴付けるために，階層クラスタ分析法を用いてキーワードのクラスタリングを行った．階層クラスタ分析とは，グラフ中のノードをいくつかのグループに分けることであるが，そのクラスタを生成する手法としては共起数の逆数を距離として，最長距離法（CLINK：complete linkage clustering method）を用いた．この手法はまず，最も距離の近いノード同士を統合しクラスタを生成し，新たに統合されたクラスタと他のクラスタとの距離にはそれぞれのクラスタに属するノードのうちで最も遠いもの同士の距離として定め，ある閾値以内にあるノードはクラスタに加えてゆく．こうしてできたそれぞれのクラスタに含まれるキーワード同士は関係が強く，それらに関する知識はすでに知られていると見なすことができる．

ミクロビューアプローチでは階層クラスタ分析法により生成されたクラスタの数を用いて，発見ルールに関しては以下のような仮定を設けた．

(1) 既知のルールのクラスタ数は 1 である．
(2) 未知のルールのクラスタ数は 2 である．
(3) ゴミのルールのクラスタ数は 3 以上である．

クラスタ数が 1 の場合は，すべてのキーワード対に関する研究が活発な場合であるので，そのようなキーワードを含むルールは既知であると考えられる．クラスタ数が 2 の場合は，キーワード集合が二つのクラスタにグループ化される場合である．それぞれのグループに関する研究は行われていても，グループ間の研究は十分ではないと考えられるので，それに関するルールは未知であると判定できる．しかし，クラスタ数が 3 以上の場合は，ゴミルールとした．これは未知の関係があるグループが 3 以上存在する場合には，ルールとして複雑すぎると考えられるからである．

例えば，クラスタリングの閾値を距離 1（キーワード間の共起数が 1 件のもの）とすると，図 7 のルールでは，全てのキーワードが 1 つのクラスタに統合さる．図 8 のルールでは GPT, G-GTP と female が一つのクラスタに統合され，hemolysis と blood group a は統合されないので，クラスタ数は 3 となる．

以上の仮説の妥当性を検討するために，図 3 で示すアンケート調査（問 1）の結果を用いた．図 9 にクラスタ数と選択肢の割合の関係をグラフにして示す．それぞれの選択肢が選ばれた割合と，階層クラスタ分析の結果のクラスタ数との関係をプロットしたものである．なお，クラスタリングの閾値として距離 1 を用いた．仮説の検証を行うために，χ^2 検定を用いて，クラスタ数の変化に伴い，それぞれの選択肢の割合に有意な変化があるかを検証した．有意水準としては 95 %を用いた．

その結果,「当然」の割合は，クラスタ数 1 と 2，1 と 3 の間に有意な差があることが認められた.「おもしろい」の割合にも，クラスタ数 1 と 2，1 と 3 の間に有意な差があることが認められた.「無意味」の割合は，クラスタ数の変化による差は認められなかった．よって，クラスタ数が 1 から複数になると「当然」の割合は減少し,「おもしろい」の割合は増加する，と言える．また,「無意味」の割合はクラスタ数の変化と関係があるとは言えない．このことから，クラスタ数が 1 のルールは「当然」なものであり，複数個あるルールは「おもしろい」ものであると考えられる．

今回の調査では「無意味」と回答する割合は少なかった．したがって，発見ルールをフィルタリングしてしまうよりは，クラスタ数の多い順にランキングする方が適切であると考えられる．

5. 文献ヒット数の経年変化

これまで述べた手法は情報検索として単純なキーワード検索手法を用いて，そのヒット数から発見ルールに関する研究動向を推測しようとしていた．Pubmed などの検索システムでは単純なキーワード検索だけでなく，文献の出版年ごとの検索も可能であり，文献ヒット数の経年変化も知ることができる．本節では文献ヒット数の経年変化が研究動向と相関があることを示す．

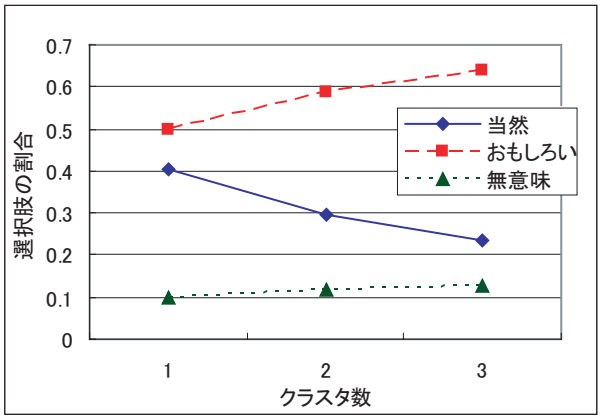

図 9 マクロビューアプローチの評価

MEDLINE に収録される文献数は年毎に増加の一途をたどり，一般的に任意のキーワードに対して文献ヒット数の経年変化は増加の傾向がある．そこで単純な文献ヒット数に代わる指標として Jaccard 係数 [村田 02] がある．属性キーワード $K_1, \cdots K_n$ に対して，その Jaccard 係数は $h(\{K_1, \cdots, K_n\})/(h(\{K_1\}) + \cdots + h(\{K_n\}))$ で与えられる．ここで $h(L)$ はキーワード集合 L で文献検索したときのヒット数である．Jaccard 係数は複数のキーワードの関連を示す一つの指標である．

Jaccard 係数の経年変化による発見ルールの分類は以下のように仮定できる．

(1) Jaccard 係数が上昇傾向にある．これはその分野の研究が盛んに行われホットな分野であることが伺える．
(2) Jaccard 係数が下降傾向にある．これはその分野の研究が収束に向かっていることが伺える．
(3) Jaccard 係数が高いまま変わらない．これは属性間の関係が常識的なものになっていることを示している．
(4) Jaccard 係数が低いまま変わらない．これはまだあまり研究されていない分野である．属性間の関係が見当はずれである場合にも生じる．

本手法の有効性を示すために，肝炎に関する五つの代表的なウイルス名（hav, hbv, hcv, hdv, hev）と hepatitis（肝炎）との間の Jaccard 係数を年毎に求め，その関係を図 10 にプロットした．また肝炎ウイルスの発見の歴史を表 1 示す [清澤 99].

図 10 と表 1 から分かるように肝炎ウイルス発見の時期と Jaccard 係数には明らかな相関がある．また肝炎研究の中で B 型肝炎 (hbv) と C 型肝炎 (hcv) が主要なものであり，C 型肝炎に関してはウイルスの発見に伴い急激に研究が進展していることが分かる．

以上のことから，発見ルールに含まれる属性名キーワード間の Jaccard 係数は，それらに関する研究の活性度を

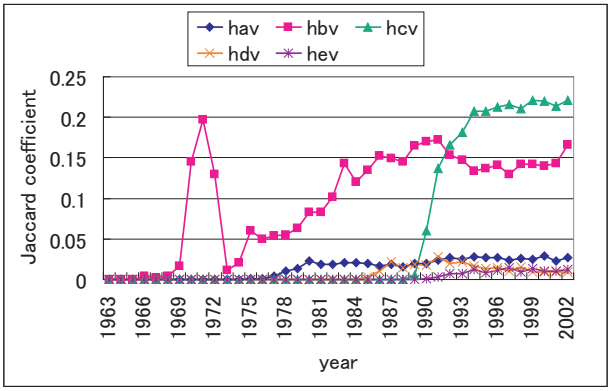

図 10 肝炎ウイルスに関する Jaccard 係数の経年変化

1965 年	オーストラリア抗原の発見.
	B 型肝炎ウイルス発見の端緒となる.
1973 年	A 型肝炎ウイルスの発見.
1977 年	デルタ抗原の発見.
	D 型ウイルス発見の端緒となる.
1983 年	E 型肝炎ウイルス粒子の同定.
1989 年	C 型肝炎ウイルス遺伝子クローニングに成功.

表 1 肝炎ウイルスの発見年表

示しており，発見ルールフィルタリングの一つの指標とすることが考えられる．

6. ま と め

情報検索手法を用いて，データマイニングにより得られた発見ルールを利用者にとって新規なものにフィルタリングする発見知識フィルタリングの試みについて述べた．その手法として発見ルールに直接関連する文献を検索するマクロビューアプローチと，発見ルールに含まれる全てのキーワード対で検索することでルールに関する研究動向を大まかに推測しようとするミクロビューアプローチを示し，医学生へのアンケート調査に基づく評価を行った．また Jaccard 係数の経年変化に基づく研究動向の推測についても示した．

現在の文献検索手法は発見ルールから抽出される属性キーワードと領域キーワードのみを用いている．したがって文献検索の精度は必ずしも高くはなく，発見ルールと直接関連しないような文献が得られる場合もある．この精度を上げることが今後の課題であるが，そのためには自然言語処理の手法を用いる必要があるであろう．

例えば，発見ルールに含まれるキーワードが文献アブストラクト中で離れた場所に存在するとするならばそれらの属性の関連を述べた文献である可能性は低いかもしれない．したがってアブストラクトの構文解析を行うことにより，属性キーワードが同一文中に現れるかどうかを確認できればフィルタリングの精度は向上すると考えられる．さらに属性キーワードが結果や結論に関連する文の中に現れるかどうか，属性キーワード間の関係を修飾する文節は何か，などということが明らかになればフィルタリングの精度はさらに向上すると考えられる [Shimbo 03]．

謝 辞

本研究は文部科学省科学研究補助金特定領域研究（課題番号：13131209）によるものである．

◇ 参 考 文 献 ◇

[北村 01] 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)

[Kitamura 02] Kitamura, Y., Park, K., Iida, A., and Tatsumi, S.: Discovered Rule Filtering Using Information Retrieval Technique. Proceedings of International Workshop on Active Mining, 80-84 (2002)

[Kitamura 03] Kitamura, Y., Iida, A., Park, K., and Tatsumi, S.: Micro View and Macro View Approaches to Discovered Rule Filtering, Proceedings of 2nd International Workshop on Active Mining, 14-21 (2003)

[Kitamura 04] Kitamura, Y., Iida, A. and Park, K.: Preliminary Evaluations of Discovered Rule Filtering Methods. Proceedings of 3rd International Workshop on Active Mining, 53-62 (2004)

[清澤 99] 清澤 研道: ウイルス肝炎とは，Medical Practice, 16(9):1394-1401 (1999)

[Motoda 02] Motoda, H. (Ed.): Active Mining: New Directions of Data Mining, IOS Press, Amsterdam (2002)

[村田 02] 村田 剛志: サーチエンジンを利用した知識発見のための視覚化，人工知能学会知識ベースシステム研究会資料, SIG-KBS-A201, 117-122 (2002)

[Ohsaki 02] Ohsaki, M., Sato, Y., Yokoi, H., and Yamaguchi, T.: A Rule Discovery Support System for Sequential Medical Data - In the Case Study of a Chronic Hepatitis Dataset -, Proceedings of International Workshop on Active Mining, 97-102 (2002)

[Onoda 02] Onoda, T., Murata, H., and Yamada, S.: Interactive Document Retrieval with Active Learning, Proceedings of International Workshop on Active Mining, 126-131 (2002)

[Shimbo 03] Shimbo, M., Yamasaki, T., and Matsumoto, Y.: Using Sectioning Information for Text Retrieval: a Case Stduy with the MEDLINE Abstracts, Proceedings of 2nd International Workshop on Active Mining, 32-41 (2003)

〔担当委員：××○○〕

19YY 年 MM 月 DD 日 受理

著 者 紹 介

北村 泰彦（正会員）

1988 年大阪大学大学院大学院博士課修了．工学博士．同年大阪市立大学工学部電気工学科助手．同情報工学科助教授を経て，2003 年関西学院大学理工学部情報科学科教授．マルチエージェントシステム，ヒューリスティック探索，WWW 情報統合の研究に従事．IEEE, AAAI, ACM, 人工知能学会，情報処理学会，ソフトウェア科学会等の会員．2001 年人工知能学会全国大会優秀論文賞，CIA-02 System Innovation Award 受賞．