社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE

# MEDLINE 情報検索に基づく発見ルールフィルタリングシステム

北村　泰彦† 　飯田　暁† 　朴　勤植‡ 　辰巳　昭治†

†大阪市立大学大学院工学研究科　〒558-8585　大阪市住吉区杉本 3-3-138
‡大阪市立大学大学院医学研究科　〒545-8585　大阪市阿倍野区旭町 1-4-3

E-mail: †{kitamura, iida, tatsumi}@kdel.info.eng.osaka-cu.ac.jp, ‡kspark@msic.med.osaka-cu.ac.jp

**あらまし**　データマイニングシステムは与えられた大量のデータに隠されている知識を自動的に発見してくれる．しかしながら発見された知識は新規なもので，また利用者にとって興味深いものとは限らない．そこで本論文では情報検索技法を用いることで発見されたルール形式の知識を新しく，利用者にとって興味のあるものにフィルタリングする手法を提案する．本手法では，インターネット上の情報源からの検索結果に応じて発見ルールをランク付けする．本論文では，医療データマイニングと MEDLINE 文献検索という具体的な事例を用いて，発見ルールフィルタリングの手順を示す．また肝炎データマイニングによる発見されたルールをフィルタリングするプロトタイプシステムを示す．最後に発見ルールフィルタリングに対する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議論する．

**キーワード**　発見ルールフィルタリング，データマイニング，情報検索，適合性フィードバック，MEDLINE データベース

# Discovered Rule Filtering System Using MEDLINE Information Retrieval

Yasuhiko KITAMURA† 　Akira IIDA† 　Keunsik PARK‡ 　and 　Shoji TATSUMI†

†Graduate School of Engineering, Osaka City University　3-3-138, Sugimoto, Sumiyoshi-ku, Osaka, 558-8585
‡Graduate School of Medicine, Osaka City University　1-4-3, Asahi-Machi, Abeno-ku, Osaka, 545-8585

E-mail: †{kitamura, iida, tatsumi}@kdel.info.eng.osaka-cu.ac.jp, ‡kspark@msic.med.osaka-cu.ac.jp

**Abstract**　A data mining system can semi-automatically discover knowledge by mining a large volume of data, but the discovered knowledge is not always novel and interesting to the user. We propose a discovered rule filtering method to filter rules discovered by a data mining system and to produce ones that are novel and interesting to the user by using information retrieval technique. In this method, we rank discovered rules according to results of information retrieval from the Internet. In this paper, firstly we show the steps of discovered rule filtering by using a concrete example of clinical data mining and MEDLINE document retrieval. Secondly, we show a prototype system to filter rules discovered by a hepatitis data mining. Lastly, we discuss micro and macro view approaches to discovered rule filtering.

**Keyword**　discovered rule filtering, data mining, information retrieval, relevance feedback, MEDLINE database

## 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，ユーザ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[1]．本稿では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[7]について述べる．

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，ユーザ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，(1)ユーザが扱いきれないほどの大量の知識が得られる，(2)ユーザにとって既知の知識が得られる，(3)ユーザにとって興味のない知識が得られる，といった問題が生じる．そこで本研究ではルール形式で得られる大量の発見知識の中から

新規でユーザにとって興味のあるもののみにフィルタリングする発見ルールフィルタリングの開発を行っている．この発見ルールフィルタリングを実現するには発見されたルールが新規かどうかを判定する必要がある．このためにわれわれはインターネット上の情報源を用いている．またユーザにとっての興味の有無を判定するためには適合性フィードバック手法を用いている．

本論文では以下，2 章においてデータマイニングと情報検索の統合とその利点に関して述べる．3 章では発見ルールフィルタリングの手順について，4 章ではそれを実装したプロトタイプシステムについて述べる．5 章ではルールフィルタリングを行う際の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議論し，6 章でまとめとする．

## 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$A_1$，$A_2$,...,$A_n$に対し，それらの関係を示す大量のデータ集合$D(\subseteq A_1 \times A_2 \times \ldots \times A_n)$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$m(D) \subseteq R=\{<A_{c1},A_{c2},\ldots,A_{cm} \rightarrow A_d>\}$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)と再現率(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$B_1,B_2,\ldots,B_m$が与えられているときに，それらを含む大量の文献集合$D' \subseteq B_1 \times B_2 \times \ldots \times B_m$からキーワードの共起数を求めることと定義できる．すなわち情報検索と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$ir(D',\{B_{k1},B_{k2},\ldots,B_{kp}\}) \in \mathbf{Int}$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，共起数の多い順にソートされた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$<A_{c1},A_{c2},\ldots,A_{cm} \rightarrow A_d>$が得られたとしよう．また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，$ir(D',\{c(A_{c1}),c(A_{c2}),\ldots,c(A_{cm})\})=k$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．これに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発見されたルールが利用者にとって興味があるかどうかに関しても評価することが可能になる．

## 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知識に対して，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

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得る．静岡大学の山口グループでは，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[2]．その一例は図 1 に示すようなものである．

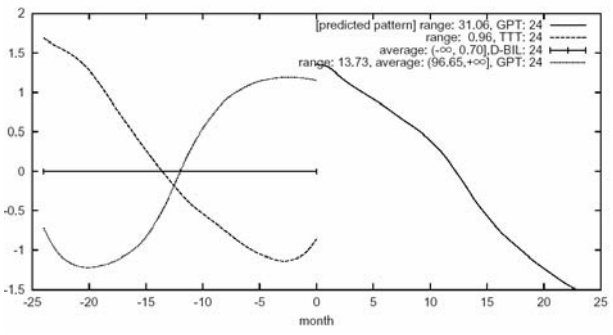



図 1 ：発見ルールの一例

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．

### 3.2. 発見ルール駆動情報検索

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 [4] (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) 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．

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，利用者の興味から導出され，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合もある．例えば，TTT は thymol turbidity test, GPT は ALT に変換する．

(2)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．

(3)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[3]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 より連続的に MEDLINE 文献検索を行う．このような文献検索を発見知識の数だけ繰り返す．当然のことながら発見知識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，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．

### 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知識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知識のソーティングを行ったり，ある閾値を設定して，それ以下のものを排除したりする．文献検索結果とソーティ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，具体的には以下のような仮説を前提とすることができる．

- 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．

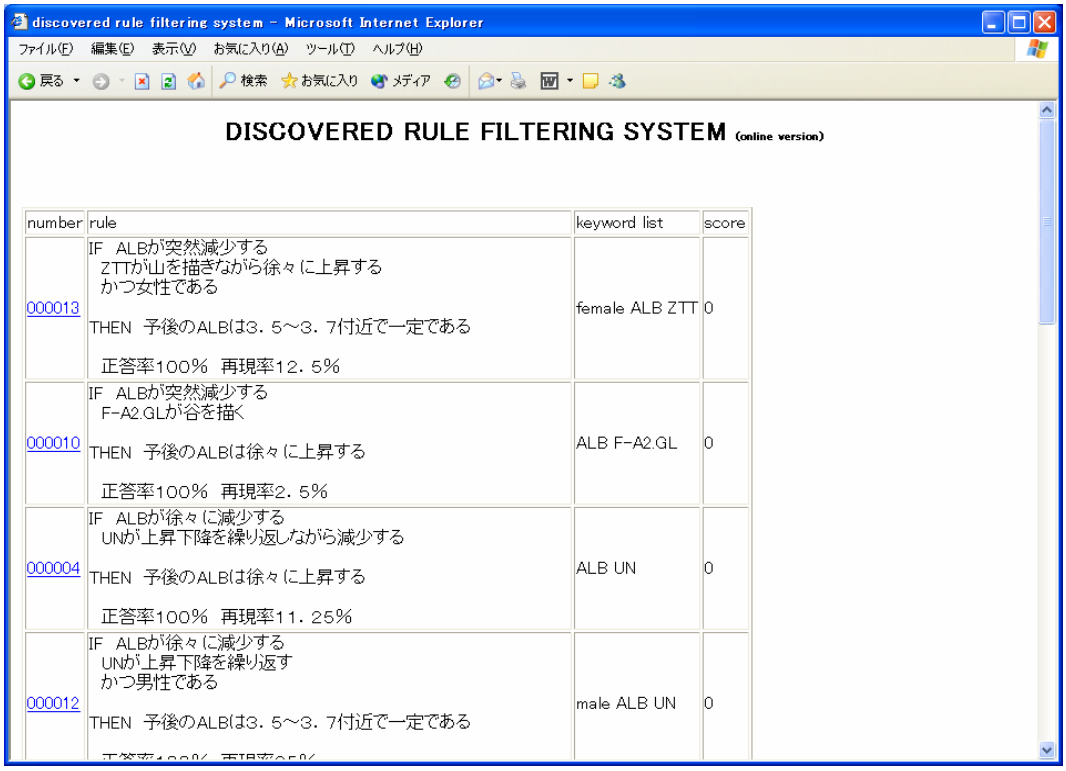
発見ルールフィルタリング手法に関しては 5 章でより詳しく述べる．

### 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，その中には発見知識と関連していても，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，ユーザ自ら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キーワードを獲得する手法として適合性フィードバック[3]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，興味あるキーワード，興味のないキーワードを自動的に抽出する．

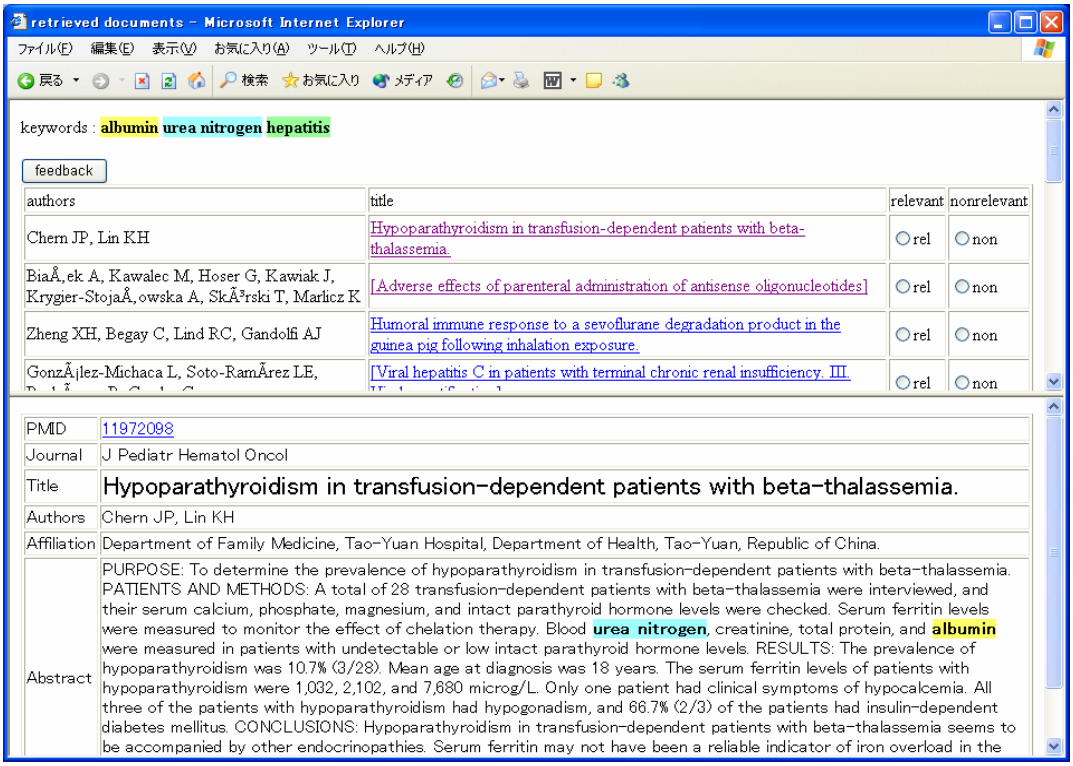
## 4. 発見ルールフィルタリングシステム

システムが起動されると図２に示すような画面が表示される．number は発見ルールの ID 番号，rule は発見ルールの記述である．keyword list には発見ルールから抽出された属性キーワードが表示されている．score はルールをランキングする際に用いる評価点である．これはこのルールと関連性のある文献数である．本システムでは未知のルールを発見しようとしているので，文献数が昇順になるように発見ルールがソートされる．ただし文献数が 0 のものは，発見ルールがゴミである可能性が高いので，ランキングを下げている．score は以下で議論する適合性フィードバックが行わ




| number | rule | keyword list | score |
|---|---|---|---|
| 000013 | IF ALBが突然減少する<br>ZTTが山を描きながら徐々に上昇する<br>かつ女性である<br>THEN 予後のALBは3．5～3．7付近で一定である<br>正答率100% 再現率12．5% | female ALB ZTT | 0 |
| 000010 | IF ALBが突然減少する<br>F-A2.GLが谷を描く<br>THEN 予後のALBは徐々に上昇する<br>正答率100% 再現率2．5% | ALB F-A2.GL | 0 |
| 000004 | IF ALBが徐々に減少する<br>UNが上昇下降を繰り返しながら減少する<br>THEN 予後のALBは徐々に上昇する<br>正答率100% 再現率11．25% | ALB UN | 0 |
| 000012 | IF ALBが徐々に減少する<br>UNが上昇下降を繰り返す<br>かつ男性である<br>THEN 予後のALBは3．5～3．7付近で一定である | male ALB UN | 0 |

図２　発見ルールフィルタリングシステムの起動画面




| authors | title | relevant | nonrelevant |
|---|---|---|---|
| Chern JP, Lin KH | Hypoparathyroidism in transfusion-dependent patients with beta-thalassemia. | rel | non |
| BiaÅ‚ek A, Kawalec M, Hoser G, Kawiak J, Krygier-StojaÅ‚owska A, SkÃ³rski T, Marlicz K | [Adverse effects of parenteral administration of antisense oligonucleotides] | rel | non |
| Zheng XH, Begay C, Lind RC, Gandolfi AJ | Humoral immune response to a sevoflurane degradation product in the guinea pig following inhalation exposure. | rel | non |
| GonzÃ¡lez-Michaca L, Soto-RamÃrez LE, | [Viral hepatitis C in patients with terminal chronic renal insufficiency. III. | rel | non |

| | |
|---|---|
| PMID | 11972098 |
| Journal | J Pediatr Hematol Oncol |
| Title | Hypoparathyroidism in transfusion-dependent patients with beta-thalassemia. |
| Authors | Chern JP, Lin KH |
| Affiliation | Department of Family Medicine, Tao-Yuan Hospital, Department of Health, Tao-Yuan, Republic of China. |
| Abstract | PURPOSE: To determine the prevalence of hypoparathyroidism in transfusion-dependent patients with beta-thalassemia. PATIENTS AND METHODS: A total of 28 transfusion-dependent patients with beta-thalassemia were interviewed, and their serum calcium, phosphate, magnesium, and intact parathyroid hormone levels were checked. Serum ferritin levels were measured to monitor the effect of chelation therapy. Blood urea nitrogen, creatinine, total protein, and albumin were measured in patients with undetectable or low intact parathyroid hormone levels. RESULTS: The prevalence of hypoparathyroidism was 10.7% (3/28). Mean age at diagnosis was 18 years. The serum ferritin levels of patients with hypoparathyroidism were 1,032, 2,102, and 7,680 microg/L. Only one patient had clinical symptoms of hypocalcemia. All three of the patients with hypoparathyroidism had hypogonadism, and 66.7% (2/3) of the patients had insulin-dependent diabetes mellitus. CONCLUSIONS: Hypoparathyroidism in transfusion-dependent patients with beta-thalassemia seems to be accompanied by other endocrinopathies. Serum ferritin may not have been a reliable indicator of iron overload in the |

図３　発見ルールフィルタリングシステムにおける適合性フィードバック

れるごとに再計算され,発見ルールは再ソートされる.

number はハイパーリンクとなっており,そこをクリックすると図３に示されるように，文献検索でヒットした文献の一覧が表示される.（ここでは 000004 を選択している.）図３でも示すように検索キーワードは発見ルールから得られたものを必要に応じて拡張したものと領域キーワード hepatitis である．文献は上部のフレームでは著者名(authors)，文献名(title)が関連性(relevant/nonrelevant)ボタンとともに表示されている．title をクリックする下部のフレームに詳細文献情報として, PubMed ID (PMID), 出典名(Journal), 文献名(Title), 著者名(Authors)，所属(Affiliation)，アブストラクト(Abstract)が表示される．アブストラクトの部分には検索キーワードがハイライトして表示される．利用者はこの文献アブストラクトを読み，自分の興味と関係があるかどうかを上部フレームの関連性ボタンをクリックすることでシステムにフィードバックをかけることができる．関連性のチェックが終わり，利用者が feedback のボタンが押すと，適合性フォードバックアルゴリズムが起動され，文献の分類が行われる．この結果は図２における発見ルールの順序に反映される．

## 5. MEDLINE 文献検索に基づく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.

### 5.1. ミクロビューアプローチ

ミクロビューアプローチは文献検索をできるだけ正確に行おうとするアプローチである．現在の手法は発見ルールから直接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文献検索を行っている．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．この精度を上げるためには二つの方法が考えられる.

(1)属性間の関係を表すキーワードの利用：発見ルールを構成する属性間の関係を表すキーワードを利用することが考えられる．例えば，肝炎データマイニングにおいて属性値の変化の周期性が重要である場合には periodicity（周期性）といったキーワードが考えられる．しかしこのようなキーワードを発見ルールから直接抽出することは難しい．そこで属性間の関係を適切に表現するキーワードをオントロジのような形であらかじめ用意しておくことが必要になるであろう.

(2)自然言語処理手法の利用：もう一つの手法は得られた文献のアブストラクトを自然言語処理の手法を用いて解析することである．例えば，発見ルールに含まれる属性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.

### 5.2. マクロビューアプローチ

マクロビューアプローチは情報検索の精度はある程度犠牲にしても属性間の関係に関する大まかな傾向を観察しようとするものである．例えば属性キーワードが共起する文献数はその属性間の関係の強さを近似的に表しているといえる.

共起文献数に代わる指標としては Jaccard 係数[5]が挙げられる．属性キーワード K1 と K2 に対して，検索キーワードを{K1},{K2},{K1,K2}としたときの文献ヒット数をそれぞれ h({K1}),h({K2}),h({K1,K2})としたとき，キーワード K1 と K2 に対する Jaccard 係数は h({K1,K2})/(h({K1})+h({K2}))で与えられる．Jaccard 係数は二つのキーワードの関連性の相対的な強さをよく表している指標である.

また MEDLINE データベースは文献情報を扱っているために出版年に応じた文献検索が可能である．したがって Jaccard 係数の年毎の変化を観測することにより発見ルールに関する属性の関連度の変化がわかる．例えば，以下のような解釈が可能かもしれない.

- 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.
- Jaccard 係数が下降傾向になる．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.
- 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.
- 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.

以上の解釈の妥当性を調べる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４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[6].

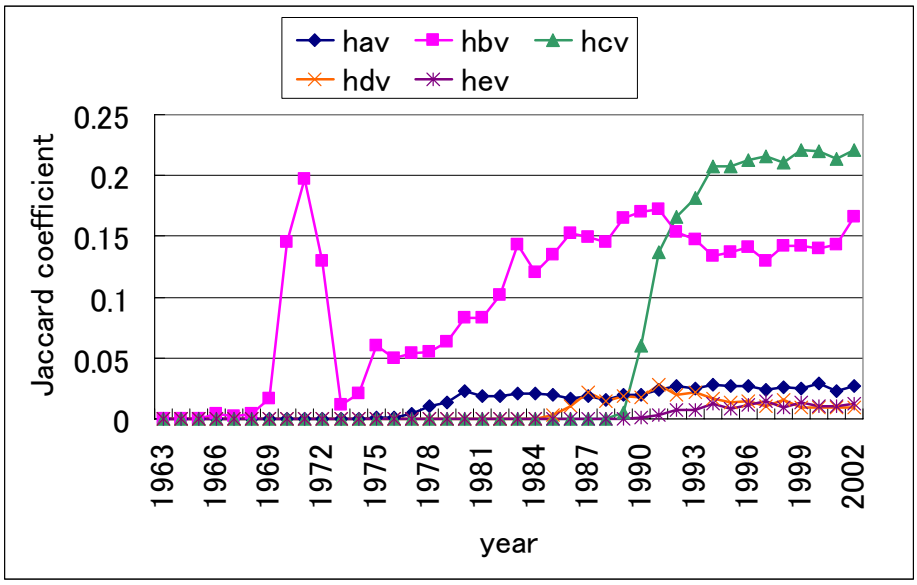



図 4　Jaccard 係数の変化

図 4 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 係数を求めることにより，それらの研究の時系列的な活性度を理解することができ，それを用いて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．

表 1　肝炎ウイルスの発見年表[6]

| | |
|---|---|
| 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． |

## 6. まとめ

MEDLINE 情報検索の結果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，利用者にとって新しくかつ興味あるものにフィルタリングする発見ルールフィルタリング手法とそのプロトタイプシステムについて述べた．今後の課題としては以下のものが挙げられる．

(1)情報検索の精度の向上．5.1 節で述べたミクロビューアプローチにより発見ルールに関連する文献検索の精度の向上を行う必要がある．また 3.5 節で述べた適合性フィードバック手法の効果も検証してゆく必要がある．

(2)発見ルールフィルタリングの効果の実証．特に，5.2 節で述べたマクロビューアプローチによりどの程度の効果があるかを実証する必要がある．

(3)発見ルールフィルタリングの応用．開発した手法を肝炎データなどのデータマイニング支援に応用し，その有効性を示す．

**謝辞：**本研究にあたり，肝炎データベースからの発見ルールを提供していただいた静岡大学山口高平教授，適合フィードバックシステムを提供していただいた国立情報学研究所の山田誠二教授ならびに電力中央研究所の小野田崇氏，自然言語処理に関して議論いただいた奈良先端大学院大学の松本裕二教授，肝炎に関する資料を提供していただいた横井英人氏に感謝の意を表します．

## 文　献

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

[2] 畑澤寛光，佐藤芳紀，山口高平，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，「情報洪水時代におけるアクティブマイニングの実現」平成 13 年度科学研究費補助金特定領域(B)研究成果報告書，pp.205-221, 2002.

[3] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.

[4] 懸俊彦，PubMed 活用マニュアル，南江堂，東京，2000.

[5] 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，”人工知能学会研究会資料，SIG-KBS-A201, pp.117-122, 2002.

[6] 清澤研道，ウイルス肝炎とは，Medical Practice, Vol.16, No.9, pp.1394-1401, 1999.

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