G-056

MEDLINE 情報検索に基づく発見ルールフィルタリング法

Discovered Rule Filtering Method Using MEDLINE Information Retrieval

飯田　暁[1]　北村　泰彦[2]　朴　勤植[3]　辰巳　昭治[1]
Akira Iida　Yasuhiko Kitamura　Keunsik Park　Shoji Tatsumi

１．はじめに

データマイニングは大量のデータ集合の中から，自動的にユーザにとって有用な知識を発見する手法である．しかし，データマイニングによって発見される知識には(1)ユーザにとって既知の知識が含まれる，(2)ユーザにとって興味のない知識が含まれる，という問題がある．そこで，我々はこれらのうち特に(1)の問題について，インターネットの情報検索結果を用いることで，ある知識が既知かどうかを判断し，新規な知識だけを取り出すためのフィルタリング手法について研究している．

本論文では，具体的な事例として，肝炎患者の検査データをマイニングすることで発見されたルールを，医学・生物学関係の文献データベースである MEDLINE（MEDlars on LINE）の文献検索結果を用いることでフィルタリングする手法について述べる．

2．データマイニングと情報検索

2.1　肝炎データマイニング

本論文では肝炎患者の検査データをデータマイニングの対象としている．このデータは千葉大学医学部より提供されたもので，データマイニングは静岡大学の山口研究室により行われている[1].

データマイニングによって得られる知識は，IF-THEN 形式のルールで表現される．ルールを図示したものの一例を図１に示す.

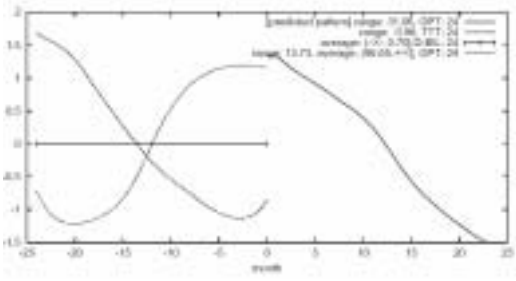

図１：発見ルールの一例

これは，「２４ヶ月間，D-BIL（直接ビリルビン）が一定で，TTT（チモール混濁試験）が減少し，GPT（グルタミン酸ピルビン酸トランスアミナーゼ）が増加するならば，その後２４ヶ月間の GPT は減少する」というルールを表している．

2.2　MEDLINE 情報検索

発見されたルールに関する情報をインターネットでの情報検索を用いて収集する．しかし，肝炎に関する情報を集めるには一般のインターネット情報源は雑多な情報が多く含まれているので，本研究では医学・生物学の文献に特化したデータベースである MEDLINE の文献検索を用いる．MEDLINE は世界各国で出版された 4000 誌を超える医学・生物学系の学術雑誌に収録された文献のデータベースで 1100 万件以上の文献が収録されている．また，MEDLINE の無料検索サービスとして PubMed (http://www4.ncbi.nlm.nih.gov/entrez/query.fcgi)が提供されている．PubMed では一般のサーチエンジンと同様にキーワードでの検索をすることができ，さらに文献が出版された日付や，研究のカテゴリーなどのオプションをつけて検索することができる．

2.3　情報検索に基づくルールフィルタリング

発見ルールフィルタリングは，データマイニングシステムにより発見されたルールのなかから，情報検索を用いることで，既知のものを取り除く．ある程度の医学的知識を持っているユーザを想定するので，医学界で常識となっているような知識は既知なものとする．また，そのような知識は MEDLINE のなかに記述されているものと仮定している．

3．ルールフィルタリングへのアプローチ

MEDLINE 情報検索を用いたルールフィルタリングを実現するための二つのアプローチを提案する．

3.1　ミクロビューアプローチ

MEDLINE データベースの中にルールの内容と同じようなことを述べている文献があれば，そのルールは既知なものだと判断することができる．そのような文献を見つけ出そうとするのが，このアプローチである．具体的には，ルールに含まれている各属性名をキーワードとして文献検索を行い，ヒットした文献の中からルールと関係するものを見つける．しかし単純なキーワード検索では，文献検索の精度は高くなく，ヒットした文献の中にルールと直接関係のあるものがほとんどない場合もある．また，ヒット数が大きくなったときに文献を一つずつ調べるには膨大なコストがかかる．これらの問題は，今後，適合性フィードバック[2]や自然言語処理などの技術を用いて改善させることが期待される．

3.2　マクロビューアプローチ

この手法は文献検索の精度はある程度犠牲にしても，属性間の関係の深さをおおまかに知ることにより，そのルールが既知なものかどうかを判断する．また，PubMed の機能としてある年代別の文献検索を用いて，属性間の関係の強さの経年変化を見ることで，それに関する研究がどの程度進んでいるかを知ることができ，それが常識となっているかどうかを判断することができる．

属性を表すキーワード間の関係の深さを表す指標としては，複数のキーワードでの共起文献数やキーワード間の Jaccard 係数などが挙げられる．共起文献数とは複数のキー

1　大阪市立大学大学院工学研究科
2　関西学院大学理工学部
3　大阪市立大学大学院医学研究科

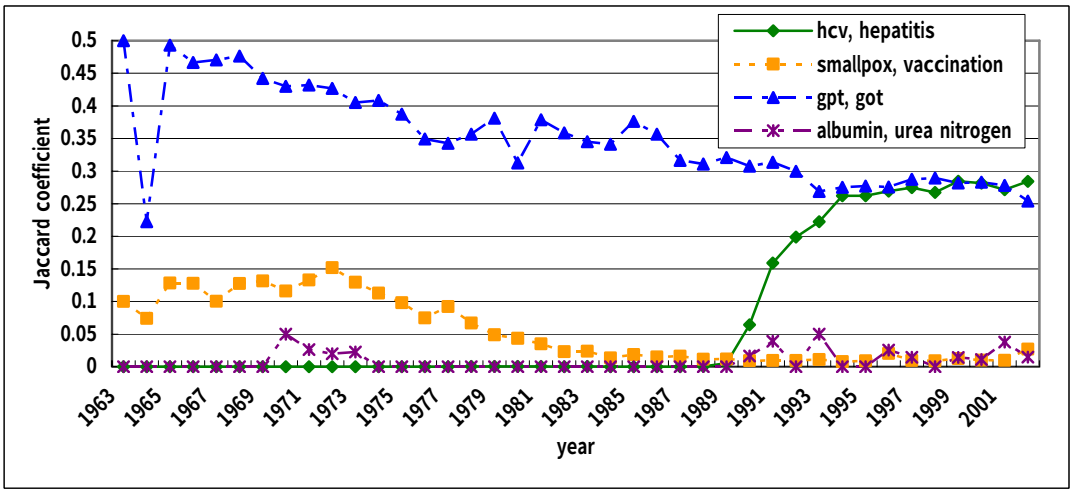

図２：キーワードの４つの組み合わせにおける Jaccard 係数の経年変化

ワードで AND 検索を行ったときのヒット数である．しかし， MEDLINE 中の文献数は年々増加する傾向にあるので，共起文献数も年毎に増加する傾向にあり，その傾向をつかむのにはふさわしくない．一方，Jaccard 係数は情報検索において，キーワード間の関係の強さを表すために用いられる係数である[3]．あるキーワード X で情報検索したときのヒット数を $n(X)$ とすると，キーワード A と B の間の Jaccard 係数は

$$Jaccard(A,B)=\frac{n(A\cap B)}{n(A\cup B)},$$

と定義される．この Jaccard 係数の経年変化を見ることで，そのキーワードに関する研究がどの程度まで進んでいるかを判断する．それには次のような仮定を用いている．

- Jaccard 係数が上昇傾向にあるとき，その分野に関する研究が盛んに行われていて，ホットな分野である．
- Jaccard 係数が下降傾向にあるとき，その分野に関する研究は既に終わっていて，収束に向かっている．その知識は既に常識的なものとなっている．
- Jaccard 係数が高いままであるとき，その関係は常識的なものになっていて，頻繁に用いられている．
- Jaccard 係数が低いままであるとき，その関係についての研究はほとんどなされておらず，常識であるとはいえない．

4．マクロビューアプローチの検証

3.1 節で述べたように，ミクロビューアプローチは問題点が多く，現段階ではうまくいかない．よって，ここからはマクロビューアプローチについて検証を行う．

3.2 節で述べた四つの仮説を検証するために，(hcv[*1], hepatitis[*2]), (smallpox[*3], vaccination[*4]), (gpt, got), (albumin[*5], urea nitrogen[*6]), の四つのキーワードの組み合わせで Jaccard 係数の経年変化を観測し，その結果を図２に示す．それぞれ４本のグラフの動きと，そのキーワードに関する医学的な事実の関係を表１に示す．

表１より Jaccard 係数の経年変化が医学的な発見や常識を表していることがわかる．

*1: C 型肝炎ウイルス　*2: 肝炎　*3: 天然痘
*4: 種痘　*5: アルブミン　*6: 尿酸窒素

表１：グラフと医学的事実の関係[4],[5]

キーワード	グラフの動き	医学的事実
hcv, hepatitis	1990 年付近から急上昇	1989 年 C 型肝炎ウイルス遺伝子クローニング成功
smallpox, vaccination	1970 年代後半ごろより下降	1977 年最後の患者発生 1980 年根絶宣言
gpt, got	高い値を維持	両者には深い関りがある
albumin, urea nitrogen	低いまま	両者の関係が注目されることはほとんどなかった

5．まとめ

データマイニングにより発見されたルールを MEDLINE からの文献検索を用いてフィルタリングする手法を提案した．その有効性を確認するために，仮説を立て，実際に Jaccard 係数の経年変化を求めて，その仮説の検証を行うことにより，それを基準として既知なルールをフィルタリングできる可能性を示した．今後は，以上のことを用いて，実際にフィルタリングを行うシステムを構築し，その有効性を検証したい．

謝辞：本研究の一部は文科省科研費特定領域研究（課題番号 13131209）によるものである．

文　献

[1] 大崎美穂，他．時系列医療データにおけるルール発見支援システムー慢性肝炎データセットでのケーススタディー．信学技報 AI，Vol. 102, No. 711，pp. 1-6, 2003
[2] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999
[3] H. Kautz, B. Selman, M. Shah. The Hidden Web. AI Magazine. Vol. 18, No. 2, pp. 27-36, 1997
[4] 清澤研道．ウイルス肝炎とは．Medical Practice．Vol.16, No.9, pp. 1394-1401，1999
[5] 国立感染症研究所感染症情報センター．感染症発生動向調査週報．
http://idsc.nih.go.jp/kansen/k01_g3/k01_40/k01_40.html
[6] Yasuhiko Kitamura, Keunsik Park, Akira Iida, and Shoji Tatsumi. Discovered Rule Filtering Using Information Retrieval Technique. Proceedings of International Workshop on Active Mining, 80-84, 2002.