文献情報検索支援システムのROC解析による相関ルール選択基準

川原 稔[†] 河野浩之[††]

文書データマイニング技術を応用して，文献情報データベースに対して，相関ルール導出アルゴリズムを適用した検索式生成支援システムの構築を行っている．しかし，相関ルール導出に用いる閾値がシステム設計者により与えられているため，必ずしも導出された相関ルールが検索式を適切に改善できるような閾値であるとは限らない．そこで，本論文では，ROC (Receiver Operating Characteristics) グラフを利用して，導出される相関ルールが閾値によってどのように変化するかを表現する．それに対してROC凸包を用いることで，より検索精度の高い相関ルールを導出する閾値の決定を行うことが可能となった．

Performance Evaluation of Bibliographic Navigation System with Association Rules from ROC Convex Full Method

MINORU KAWAHARA[†] and HIROYUKI KAWANO[††]

We have been constructing bibliographic navigators using association rules. In order to provide effective knowledge for naive users, it is very important to decide several threshold values for mining rules. In this paper, we focus on the techniques of ROC graph to evaluate the characteristics of derived rules. By using the ROC convex full method, we can estimate appropriate threshold values to derive association rules for keywords.

1. まえがき

図書・文献データベースに対する情報検索では，一般に検索領域に対する領域知識に加えて，検索システムに習熟することが必要であるため，目的のデータを得るのが難しい[9]．スムーズに検索を行うために熟練した図書館司書の支援に頼ることも多い．このような煩雑さを解消あるいは緩和するために，情報検索システム構築にかかわる研究が数多く行われている[2),3),5),7),9),11)]．そこで，我々は，データマイニング手法[6),8)]の一つである相関ルール (association rule) 導出アルゴリズム[12)]を拡張して，文献情報検索に適用した支援システムのプロトタイプを開発して実証実験を行っている[2),3),5)]．そのシステムは，**図1**のように，導出されたルールにもとづいた関連キーワードを検索ユーザに提示することで，検索にかかわる知識を与えて検索支援を行うものである．

現在のシステムでは，複数属性の関係にもとづいたアルゴリズム[2),3)]を用いて，導出にかかわる最小サポート閾値 *Minsup* 及び最小確信度閾値 *Minconf* のうち，*Minsup* を次のように動的に決定している．関連キーワードが導出されない場合は，タイトル属性と著者属性を用いて拡大したキーワード空間から導出される関連キーワード集合と，アブストラクト属性から導出される関連キーワード集合の空でない共通部分が得られるまで，*Minsup* をシステムの設定限界値まで段階的に緩和して関連キーワードを導出している．逆に，導出される関連キーワードがシステムの設定値 *Maxkey* より多く導出された場合には，導出される関連キーワード数が *Maxkey* 以下になるまで，*Minsup* を厳しくして関連キーワードの導出数を抑制している．これらの閾値はシステム管理者により与えられるものであるが，導出される関連キーワードを選択するための指針，すなわち，相関ルール選択基準を妥当に決定する方法が文献情報検索支援システムにおいて必要である．そこで，本論文では，様々な分野でパフォーマンス空間を解析するのに有効である ROC (Receiver Operating Characteristic) 解析手法[1),10)] を用いて，検索要求に対する相関ルール導出にかかわる閾値を決定する方法について議論し，実装システムによる実際のデータを用いての評価を行う[4)]．

以下，2章で，ROC解析手法についての概略を述

† 京都大学大型計算機センター
Data Processing Center, Kyoto University
†† 京都大学大学院情報学研究科
Department of Systems Science, Kyoto University

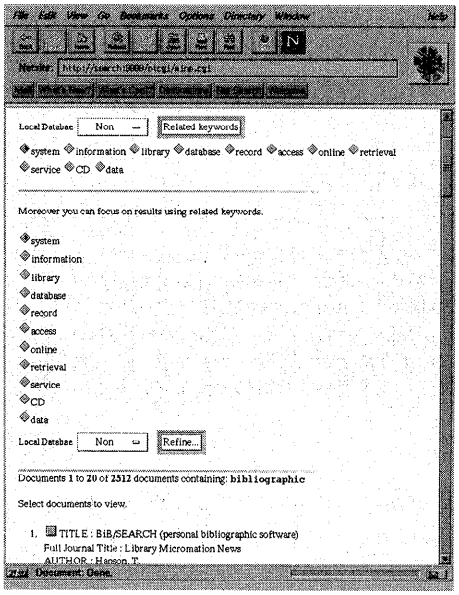


図 1 文献情報検索支援システムの検索結果画面
Fig. 1 The result window of our system.

べる．3 章では，ROC 解析手法を文献情報検索支援システムに適用する際に必要となる各種パラメータを定義し，適用方法を示す．4 章では，実装システムによるデータにもとづいたパラメータを設定し，ROC 解析による閾値の設定に対する効果を評価する．また，5 章において，結論と将来の課題について述べる．

2. ROC 解析手法

ROC グラフは，分類子のパフォーマンスをクラス分布やコスト分布から分離して視覚化することにより，クラス分布やコスト分布を正確に把握することが困難な場合でも，分類子のパフォーマンスを比較することを可能にする手法である．ROC 凸包は，ROC グラフに対して解析的手法を適用して，最大のパフォーマンスをもつ分類子を決定する手法である[1),10)]．

2.1 ROC グラフ

ある事象が 2 つの事象クラス “正の事象クラス: P (positive)” と “負の事象クラス: N (negative)” に分類でき，その事象に対する分類子による分類を，“正: y (yes)” と “負: n (no)” とする．このとき，正の事象 P が正 y と正しく分類される比率 TP は，事象 P が分類 y となる事後確率 $p(y \mid P)$ により，

$$TP = p(y \mid P) \simeq \frac{\text{正であると分類された正の事象}}{\text{すべての正の事象}} \quad (1)$$

と表すことができ，負の事象 N が誤って正 y と分類される比率 FP は，事象 N が分類 y となる事後確率 $p(y \mid N)$ により，

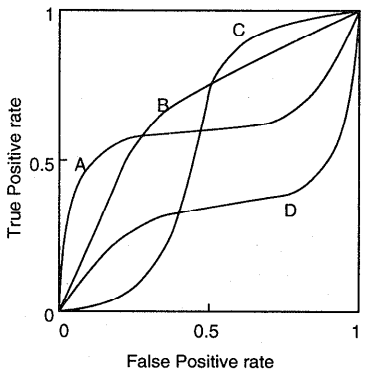


図 2 4 つの分類子による ROC グラフ
Fig. 2 An ROC graph of four classifiers

$$FP = p(y \mid N) \simeq \frac{\text{正であると分類された負の事象}}{\text{すべての負の事象}} \quad (2)$$

と表すことができる．

いくつかの事象 I に対して，FP を X 軸の値，TP を Y 軸の値としてプロットすると **図 2** のような ROC カーブと呼ばれるグラフが描かれ，これを分類子のパフォーマンスを表すのに用いる．ROC グラフでは，グラフが上端に近づくほど，すなわち TP がより高くなるほど，分類子により事象が正確に分類されたことになる．逆に，グラフが右端に近づくほど，すなわち，FP がより高くなるほど，分類子による分類にノイズが入ってくることになる．したがって，TP がより高く FP がより低い点の方，つまり左上端に ROC グラフが近づくように描かれるほど，よりパフォーマンスが高いといえる．図 2 は 4 つの異なる分類子のパフォーマンスを表しているが，例えば，分類子 A による ROC カーブは分類子 D による ROC カーブより常に左上に存在しているので，分類子 A の方がよりパフォーマンスが高いことになる．

2.2 ROC 凸包

ROC グラフでは，事象クラスやコストを切り放して視覚化することによりパフォーマンスを表しているため，コストを考慮した解析も必要である．ここで，c(分類, 事象クラス) を “分類” 及び “事象クラス” の 2 次のエラーコスト関数とすると，正の事象クラスを負に分類したときのエラーコストは $c(n \mid P)$，負の事象を正に分類したときのエラーコストは $c(p \mid N)$ と表せる．また，正の事象の事前確率を $p(P)$ とすると，負の事象の事前確率は $p(N) = 1 - p(P)$ となる．よって， ROC グラフ上の点 (FP, TP) に対する分類子のコストは，

$$p(P)\cdot(1-TP)\cdot c(n,P)+p(N)\cdot FP\cdot c(y,N) \tag{3}$$

により表されることになる．

ここで，ROC グラフにおける 2 点 (FP_1,TP_1) 及び (FP_2,TP_2) を考えると，これら 2 点のコストが等価であるとき，

$$\frac{TP_2-TP_1}{FP_2-FP_1}=\frac{p(N)\cdot c(y,N)}{p(P)\cdot c(n,P)} \tag{4}$$

となる．この等式は，コストの等価な 2 点 (FP_1,TP_1) と (FP_2,TP_2) を通る ROC グラフ上の等パフォーマンス線 (iso-performance line) の傾きを与えている．したがって，等パフォーマンス線の傾きは，$p(N)/p(P)$ とエラーコスト比 $c(y,N)/c(n,P)$ により決定される．例えば，負の事象が発生する確率と正の事象が発生する確率の比が 3 : 1 $(p(N)/p(P)=3)$ である場合に，負の事象が発生しているにもかかわらず y と分類されてしまうコストに対して，正の事象が発生しているにもかかわらず n と分類されてしまうコストが等しいときの等パフォーマンス線の傾きは 3 となり，10 倍コストがかかるときの傾きは 3/10 となる．

そして，この傾きの直線を最も左上端の点 $(0,1)$ に近い位置に描くことができる分類子が最も高いパフォーマンスを示し，各分類子による ROC カーブに直線が接するように凸包状に境界を形成した ROC 凸包が**図 3** の陰をつけた部分である．例えば，図 3 において，αで示される傾き 3 の直線が接する ROC 凸包を形成する分類子は A であり，A が最も高いパフォーマンスを示す分類子である．また，βで示される傾き 1/3 の直線が接する ROC 凸包を形成する分類子は C であり，C が最も高いパフォーマンスを示す分類子である．なお，分類子 B により描かれる ROC カーブは，A 及び C による ROC カーブよりそれぞれ高いパフォーマンスを示す部分があるが，A 及び C による ROC カーブとこれら両カーブに接する等パフォーマンス線により形成される ROC 凸包に完全に覆われており，B による ROC カーブのすべての接線に対して，同じ傾きをもつ凸包に接する接線が存在する．つまり，どのような等パフォーマンス線の傾きに対しても，少なくとも A あるいは C のいずれかが B よりも高いパフォーマンスを示しているため，B が分類子として使用されることはない．

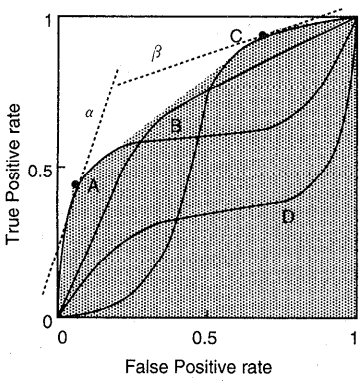


図 3 ROC 凸包における等パフォーマンス線
Fig. 3 Lines α and β show the optimal classifier under different sets of conditions.

3. 文献情報検索支援システムへの ROC 解析手法の適用

有効な閾値の決定による関連キーワードの導出を効果的に行うため，文献情報検索支援システムに ROC 解析手法を適用できるよう，$\bigcup$を集合の論理和，$\bigcap$を集合の論理積，$||$ を集合内のアイテム数を求める演算子として次を定義する．

[定義]

$\mathbf{G}$	:	検索要求キーワード集合
n	:	$\mathbf{G}$ の検索要求キーワード数
k_i	:	$\mathbf{G}$ の i 番目の検索キーワード $(1\le i\le n)$
$\mathbf{K}_i$	:	k_iが被覆する文献の集合
$\mathbf{B}$	:	$\mathbf{G}$ により被覆される文献の集合
m	:	$\mathbf{G}$ から導出されるキーワード数
r_j	:	$\mathbf{G}$ から導出される j 番目のキーワード $(1\le j\le m)$
$\mathbf{R}_j$	:	r_jが被覆する文献の集合

図 4 は，検索対象となる文献データすべての集合 $\mathbf{U}$ に対する $\mathbf{B}$ 及び $\mathbf{R}_j$による被覆状態を示している．絞込み検索においては，$\mathbf{G}$ のすべてのキーワードにより被覆される文献集合

$$\mathbf{B}=\bigcap_{i=1}^{n}\mathbf{K}_i \tag{5}$$

を絞込む事象 $\mathbf{B}$ が正となり，逆に拡大する事象$\overline{\mathbf{B}}$が負となる．したがって，正の事象でありかつ正と分類される事象は $\mathbf{B}\cap\bigcup_{j=1}^{m}\mathbf{R}_j$ であり，負の事象でありかつ正と分類される事象は$\overline{\mathbf{B}}\cap\bigcup_{j=1}^{m}\mathbf{R}_j$ であるから，TP は，

$$TP=\frac{|\,\mathbf{B}\cap\bigcup_{j=1}^{m}\mathbf{R}_j\,|}{|\,\mathbf{B}\,|} \tag{6}$$

で表すことができ，FP は，

$$FP=\frac{|\,\overline{\mathbf{B}}\cap\bigcup_{j=1}^{m}\mathbf{R}_j\,|}{|\,\overline{\mathbf{B}}\,|} \tag{7}$$

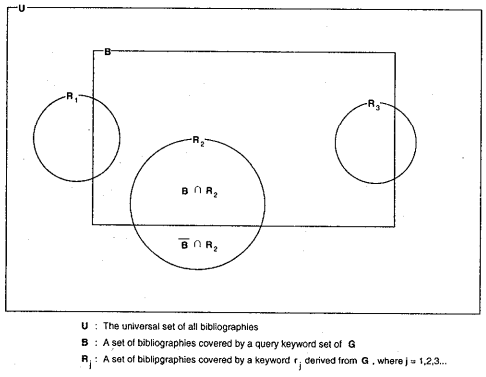


図 4 文献情報空間におけるキーワードによる被覆状態
Fig. 4 State of bibliographies covered by keywords in a bibliographic space.

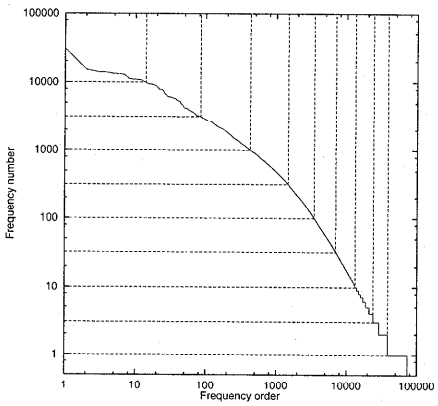


図 5 キーワード出現頻度
Fig. 5 The frequency of the commonly used words in titles.

で表すことができる．例えば，図 4 で，正の事象でありかつ正と分類される事象が $\mathbf{B} \cap \mathbf{R}_2$ であり，負の事象であるが正と分類される事象が $\overline{\mathbf{B}} \cap \mathbf{R}_2$ である．

この TP と FP をもとに ROC グラフを作成し，$Minsup$ を分類子として値を変化させたときの (FP,TP) を ROC グラフ上にプロットする．それに対して，ROC 凸包[10]を用いて等パフォーマンス線を描き，コストクラスに応じた分類子，すなわち，$Minsup$ を求める．

4. 性能評価

文献情報検索支援の実験システムには，1987 年 1 月から 1997 年 12 月の 11 年間に INSPEC☆により配布された 3,012,864 件の文献データを格納している．これらのデータをもとにして，相関ルールを導出して関連キーワードによる絞込み検索の支援を行っている．本章では，このデータをもとにして導出される関連キーワードを用いて，1997 年の 1 月から 12 月の 1 年間に INSPEC により配布された 330,562 件の文献データを分析対象として扱う．したがって，検索対象となる全文献データ数 $|\mathbf{U}|$ は 330,562 である．

これらの文献データに対して，タイトル部分で使用されているキーワードを調べると頻出順位と出現回数の関係は**図 5** のようになっており，使用されているキーワードに大きな偏りが見られた．なお，キーワードとして意味の無い “and” や “the” などの無意味語は辞書を用いて除去している．

そこで，ROC グラフ上にカーブが描かれるように，

☆ INSPEC データベースは，英国 IEE からの独立組織である INSPEC が，文献の収集・整理を行い全世界に配布している理工学系の文献二次情報であり，計算機・制御・情報工学，電子・電気工学，物理学の分野における文献データベースである．

表 1 検索キーワードのカテゴリ
Table 1 Categories of retrieval keywords.

カテゴリ	出現回数	キーワード数	サンプル数
1	10001 ～	13	13
2	3163 ～ 10000	70	20
3	1001 ～ 3162	341	20
4	317 ～ 1000	1048	20
5	101 ～ 316	1974	20
6	33 ～ 100	3562	36
7	11 ～ 32	6302	64
8	4 ～ 10	10722	108
9	2 ～ 3	14540	146
10	1	34738	348

出現回数に応じて図 5 の対数縦軸上でほぼ均等に分れるように，キーワードを**表 1** のようなカテゴリにクラス分けして，それぞれのカテゴリのキーワードによる (FP,TP) を求めて，その平均値をプロットする．表 1 の各カテゴリは，ちょうど図 5 において左上から右下に向かって順に囲まれたそれぞれの部分に相当する．各カテゴリからは，カテゴリに含まれるキーワード数の 1%以上のキーワードをサンプリングするように，表 1 の “サンプル数” 欄に示した数のキーワードを無作為抽出して評価を行った．

閾値の相互作用によるゆらぎの影響を避けるため $Minconf$ は 0.01 に固定し，$Minsup$ を変化させて，サンプリングしたキーワードから導出されるキーワードの数を求めると，全カテゴリの平均導出キーワード数は**図 6** のようになり，$Minsup$ の値が小さくなると急激に導出されるキーワード数が増加することがわかる．更に，カテゴリによる導出キーワード数を調べると**図 7** のようになり，出現頻度が高いキーワードほど導出されるキーワードのサポート値が小さくなるため，$Minsup$ を固定してしまったのでは，導出キーワード

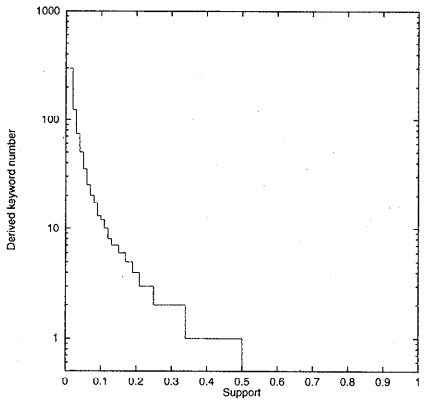


図 6 平均導出キーワード数
Fig. 6 Average number of derived keywords.

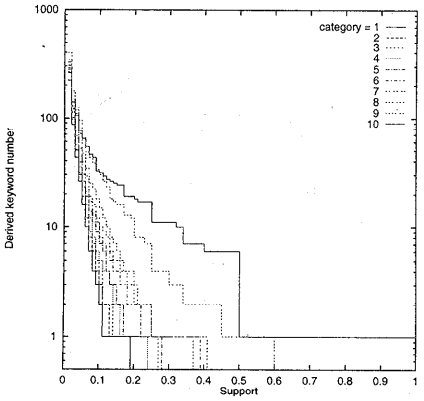


図 7 カテゴリごとの平均導出キーワード数
Fig. 7 Average number of derived keywords for each category.

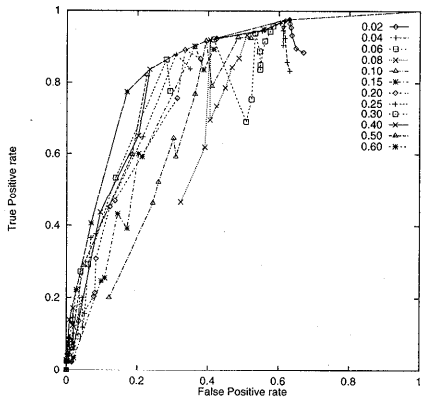


図 8 *Minsup* を分類子とした ROC グラフ
Fig. 8 The ROC graph of *Minsup* with the iso-performance line.

表 2 ROC 凸包による最適 *Minsup*
Table 2 The most suitable *Minsup* derived from the ROC convex hull.

適応範囲	分類子 (*Minsup*)
0.0000 ~ 0.0638	AllPos
0.0638 ~ 0.2597	0.02
0.2597 ~ 0.2746	0.08
0.2746 ~ 0.4126	0.10
0.4126 ~ 0.4535	0.20
0.4535 ~ 0.5478	0.25
0.5478 ~ 0.5751	0.30
0.5751 ~ 0.9644	0.40
0.9644 ~ 3.6601	0.60
3.6601 ~ 4.4116	0.60
4.4116 ~ 12.906	0.40
12.906 ~ 227.06	0.60
227.06 ~	AllNeg

数に大きなばらつきが生じてしまうことがわかる．

Minsup を分類子とした ROC カーブを描くため，図 7 において各カテゴリにおける導出キーワード数が 0 となるサポート値の近傍を取り出し，また，すべてのカテゴリにおいてキーワードが導出される 0.2 未満のサポート値に対して適宜刻みをとり，

$$Minsup = \{0.02, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2, 0.25, 0.30, 0.4, 0.5, 0.6\}$$

について評価を行った．

各 *Minsup* の値に対する各カテゴリの (FP, TP) を求めて，ROC グラフにプロットしたものが**図 8** であり，ROC 凸包の境界線も同時に描いている．図 8 における等パフォーマンス線の傾きに対する最適分類子，すなわち，*Minsup* の値を，ROC 凸包を用いて求めると**表 2** のような結果となった．なお，表 2 に示されている “AllPos” はすべての関連キーワードを導出することを表し，“AllNeg” は関連キーワードを何も導出しないことを表している．

図 8 を見ると，$TP = 1$(上端) に近い部分の凸包の境界線が，$TP = 1$ の方に突出した一部のプロットにより盛り上がってしまい，いびつな形となってしまった．また，表 2 を見ると，*Minsup* の値が 0.02 から 0.08 に急激に飛んでいることがわかる．これらの原因を調べるため，カテゴリと FP 値の関係を示したものが**図 9**，カテゴリと TP 値の関係を示したものが**図 10** である．図 9 及び図 10 を見ると，カテゴリ 8 ~10 の部分において，FP 値及び TP 値が急激に収束しているのがわかる．表 1 から，これらのカテゴリは出現回数が 10 回以下のキーワードを含むカテゴリであり，この程度の出現回数では有効な相関ルールが導出されていないことが原因と予測される．

そこで，キーワードの出現回数が 10 回以下であるカテゴリ 8~10 を含めない形で，ROC グラフを描いたものが**図 11**，ROC 凸包により *Minsup* を求めた

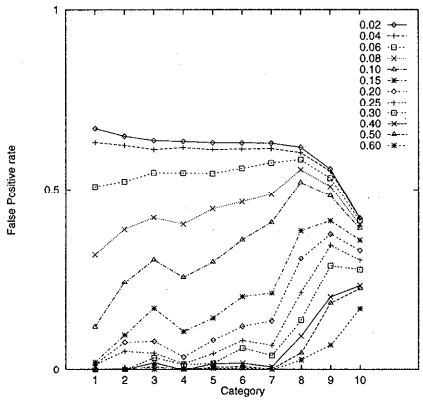


図 9 カテゴリに対する FP 値
Fig. 9 Value of FP for each category.

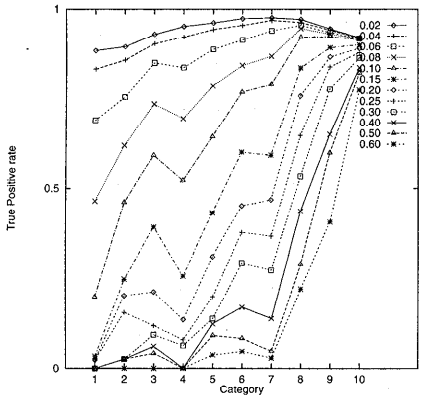


図 10 カテゴリに対する TP 値
Fig. 10 Value of TP for each category.

表 3 カテゴリ 8~10 を除外した ROC 凸包による最適 $Minsup$
Table 3 The most suitable $Minsup$ derived from the ROC convex hull without categories from 8 to 10.

適応範囲	分類子 ($Minsup$)
0.0000 ~ 0.0638	AllPos
0.0638 ~ 0.4791	0.02
0.4791 ~ 0.7195	0.04
0.7195 ~ 0.8103	0.06
0.8103 ~ 1.0589	0.10
1.0589 ~ 1.7351	0.15
1.7351 ~ 3.2032	0.25
3.2032 ~ 4.3497	0.30
4.3497 ~ 12.906	0.40
12.906 ~ 227.06	0.60
227.06 ~	AllNeg

ものが**表 3** である．

ここで，各キーワードの $p(N)/p(P)$ は，検索キーワードが被覆する文献集合を $\mathbf{B}$ とすると，

$$\frac{p(N)}{p(P)} = \frac{|\mathbf{U}| - |\mathbf{B}|}{|\mathbf{B}|} \tag{8}$$

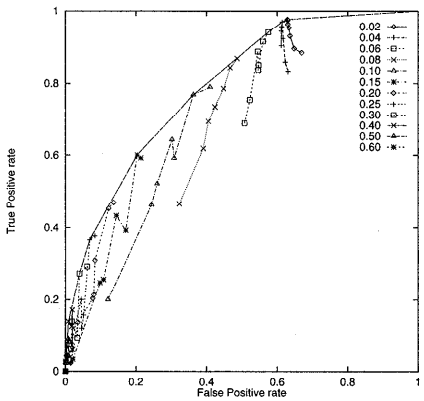


図 11 カテゴリ 8~10 を除外した $Minsup$ を分類子とした ROC グラフ
Fig. 11 The ROC graph of $Minsup$ with the iso-performance line without categories from 8 to 10.

表 4 $R_{error} = 145$ における $Minsup$
Table 4 $Minsup$ at $R_{error} = 145$.

カテゴリ	$p(N)/p(P)$ × コスト比	最適 $Minsup$
1	0.0000 ~ 0.2211	AllPos ~ 0.02
2	0.2211 ~ 0.7139	0.02 ~ 0.04
3	0.7141 ~ 2.2706	0.04 ~ 0.25
4	2.2728 ~ 7.1847	0.25 ~ 0.40
5	7.2075 ~ 22.565	0.40 ~ 0.60
6	22.790 ~ 69.076	0.60
7	71.235 ~ 207.24	0.60
8	227.97 ~ 569.93	AllNeg
9	759.91 ~ 1139.9	AllNeg
10	2279.7	AllNeg

で表すことができる．$|\mathbf{B}|$ は，キーワード検索に対するヒット件数として求まり，$|\mathbf{U}|$ は 330,562 である．

また，エラーコスト比 $c(y,N)/c(n,P)$ は，システムの管理者や検索ユーザにより指定されてもよいものではあるが，ここでは，カテゴリ 8 以下において関連キーワードが導出されない値を与えて評価を行う．表 3 から，キーワード導出が行われない AllNeg となるのは，式 (4) の値が 227.06 以上のときであるから，

$$\frac{|\mathbf{U}| - |\mathbf{B}|}{|\mathbf{B}|} \cdot \frac{c(y,N)}{c(n,P)} = 227.06$$

から，

$$R_{error} = \frac{c(n,P)}{c(y,N)} = \frac{330562 - |\mathbf{B}|}{227.06 \times |\mathbf{B}|}$$

となる．なお，R_{error} はノイズに対する検索漏れのコスト比となるので，この値を大きくとると検索漏れが減少する関係となる．

R_{error} に，$|\mathbf{B}| = 10$ を代入すると，$R_{error} = 145$，となり，この値を用いて $Minsup$ を求めたものが，**表 4** である．それらをもとに，導出されるキーワード数，キーワードが導出されない比率，$Minsup$ の値，FP 及

表 5 $R_{error} = 145$ における導出結果
Table 5 The derived result at $R_{error} = 145$.

検索カテゴリ	平均導出キーワード数											非導出率 (平均)[%]	$Minsup$ (平均)	FP (平均)	TP (平均)
	1	2	3	4	5	6	7	8	9	10	全体				
1	10	35	33	9	1	0	0	0	0	0	89	0	0.02	0.6643	0.8863
2	9	24	18	6	1	0	0	0	0	0	58	0	0.03	0.6306	0.8996
3	1	1	0	0	0	0	0	0	0	0	4	35	0.17	0.1584	0.3213
4	0	0	0	0	0	0	0	0	0	0	2	80	0.34	0.0202	0.1072
5	0	0	0	0	0	0	0	0	0	0	2	95	0.50	0.0030	0.0329
6	0	0	0	0	0	0	0	0	0	0	2	94	0.60	0.0043	0.0472
7	0	0	0	0	0	0	0	0	0	0	1	94	0.60	0.0008	0.0431
8	0	0	0	0	0	0	0	0	0	0	0	100	AllNeg	0.0000	0.0000
9	0	0	0	0	0	0	0	0	0	0	0	100	AllNeg	0.0000	0.0000
10	0	0	0	0	0	0	0	0	0	0	0	100	AllNeg	0.0000	0.0000

(注) 平均による丸め誤差のため，カテゴリ毎の平均の合計が全体の平均に一致しない場合がある．

表 6 従来手法による導出結果
Table 6 The derived result by our basic algorithm.

検索カテゴリ	平均導出キーワード数											非導出率 (平均)[%]	$Minsup$ (平均)	FP (平均)	TP (平均)
	1	2	3	4	5	6	7	8	9	10	全体				
1	2	2	0	0	0	0	0	0	0	0	5	0	0.08	0.3590	0.5130
2	2	3	1	0	0	0	0	0	0	0	6	0	0.08	0.3882	0.6519
3	2	3	1	1	0	0	0	0	0	0	8	0	0.08	0.4144	0.6990
4	2	4	2	0	0	0	0	0	0	0	9	0	0.08	0.4314	0.7386
5	3	4	2	1	0	0	0	0	0	0	10	0	0.08	0.4496	0.8260
6	2	4	3	2	1	0	0	0	0	0	12	0	0.09	0.4290	0.8123
7	2	3	3	2	1	1	0	0	0	0	12	0	0.10	0.4117	0.8409
8	1	2	2	2	1	0	0	0	0	0	8	4	0.18	0.2676	0.7040
9	1	1	1	1	1	1	0	0	0	0	8	13	0.31	0.1829	0.6355
10	1	1	2	2	1	1	0	0	0	0	10	3	0.27	0.2332	0.8453

(注) 平均による丸め誤差のため，カテゴリ毎の平均の合計が全体の平均に一致しない場合がある．

び TP の値について，カテゴリごとの平均を求めたものが，**表 5** である．また，我々の実験システムで用いているアルゴリズム[2),3)]による同様の結果を**表 6** に示す．なお，このアルゴリズムでは固定された最小サポート閾値 $Minsup = 0.08$ を用いてキーワード導出を行っているが，導出されるキーワード数が $Maxkey = 15$ を越える場合には，それを越えないように最小サポート閾値を厳しくする手法を採用しているため，表 6 では平均 $Minsup$ が 0.08 より大きくなり導出キーワード数が抑えられている部分がある．

表 5 と表 6 を比べると，表 6 では検索キーワードの出現頻度が高いものほど導出キーワード数が少なくなっているのに対して，表 5 では，検索キーワードの出現頻度が高いものほど導出されるキーワード数が多くなっていることがわかる．これは次の理由によると考えられる．

出現頻度の高い検索キーワードによる相関ルール導出では，導出対象となる文献数が多くなり相対的に導出されるルールのサポート値が小さくなってしまう．従来手法では最小サポート閾値を固定的に用いているため，検索キーワードと共起性が強く，かつ出現頻度の高いキーワードしかサポート値がその閾値を越えることができないため，結果として導出キーワード数が少なくなる．それに対して本手法では，検索キーワードの出現頻度が高くなると，$p(N)/p(P)$ が小さくなりその結果 $p(N)/p(P) \times 1/R_{error}$ が小さくなるため，表 4 の最適 $Minsup$ に見られるように相関ルール導出で用いられる最小サポート閾値が引き下げられるので，サポート値がその閾値を越えるキーワードが多くなるからである．

また，表 5 と表 6 の高頻出 (上位カテゴリ) のキーワードからの導出キーワードのカテゴリ分布を比べると，従来手法では高頻出 (上位カテゴリ) のキーワードが導出される傾向が見られるのに対して，本手法では非高頻出のキーワードも導出されている．例えば，カテゴリ 2 に属す出現頻度 5,536 のキーワード "performance" から導出されるキーワードは，従来手法では最小サポート閾値 0.08 に対して，

system(1), high(1), simulation(2), model(1),
control(1), evaluation(2), analysis(1),

network(1)

の8キーワードが導出された．各キーワードに付した括弧内の数字は，そのキーワードが属すカテゴリであるが，いずれもカテゴリ1あるいはカテゴリ2の高頻出のキーワードが導出されていた．本手法では最小サポート閾値が0.02となり，これら導出キーワードに加えて，

time(2), management(3), design(2),
computer(3), assessment(3), machine(3),
data(2), process(2), method(1), based(1),
algorithm(2), parallel(2), effect(1),
processing(2), optical(2), broadband(4)

などの他合計71キーワードが導出された．このとき，カテゴリ毎の導出キーワード数は，カテゴリ1が10，カテゴリ2が24，カテゴリ3が29，カテゴリ4が8であり，カテゴリ3やカテゴリ4に属す非高頻出キーワードも導出されていた．逆に，カテゴリ8に属す出現頻度6のキーワード“replanning”から導出されるキーワードは，本手法では最小サポート閾値はR_{error}をキーワード導出が行われない値に設定したためAll-Posとなりキーワード導出が行われないが，従来手法では動的に引き上げられた最小サポート閾値0.17に対して，

planning(3), time(2), real(3), system(1),
assembly(4)

の5つのキーワードが導出された．

ここで，表5と表6の平均FP及び平均TPからROCグラフ上で最もパフォーマンスの低い点$(1,0)$からの距離を求めたものが**表7**である．ROCグラフ上では$(1,0)$からの距離が長いほどパフォーマンスが高くなるが，表7ではカテゴリ1及び2において本手法のパフォーマンスが高いことがわかる．カテゴリ3〜7においては，パフォーマンスは両手法でほぼ同等といえる．したがって，従来手法では，高頻出キーワードから導出を行うと，その結果として同様に高頻出キーワードしか得られず[2]，それらを提示されても知識としての有効性はあまり高くないが，本手法では非高頻出キーワードも導出されるため，閾値の上限を切ることによる高頻出キーワードの導出を抑制する手法と組合せることで，パフォーマンスを保証した有効なルール導出を行うことが考えられる．

表7 ROCグラフ上の$(1,0)$からの平均距離
Table 7 The average distances from the point $(1,0)$ on the ROC graph.

カテゴリ	本手法	従来手法	本手法 − 従来手法
1	0.9477	0.8210	0.1267
2	0.9725	0.8940	0.0785
3	0.9008	0.9119	-0.0111
4	0.9856	0.9321	0.0535
5	0.9975	0.9926	0.0049
6	0.9968	0.9929	0.0039
7	1.0001	1.0263	-0.0262
8	1.0000	1.0159	-0.0159
9	1.0000	1.0351	-0.0351
10	1.0000	1.1413	-0.1413

また，表5のカテゴリ2と3の非導出率が0%から立ち上がっている部分を境に，上位のカテゴリと下位のカテゴリでは，導出キーワード数が大きく異なっている．これは，図5に見られるように，キーワードの出現頻度が100位あたりを境に大きく異なっているため，$p(N)/p(P)$の値が影響を受けているためと考えられる．したがって，非高頻出キーワードからの導出では，本手法ではキーワードが導出され難くなってしまうが，出現頻度に応じた検索漏れ及びノイズに対するパラメータをR_{error}に取り入れることで，あらゆる出現頻度のキーワードに対応したパフォーマンスを保証できる閾値の設定が可能と考えられる．

5. 結　論

現在，相関ルール導出アルゴリズムが広く用いられるようになってきたが，システム管理者が設定する閾値をいかに決定するかが問題である．本論文では，閾値を有効に決定する手法として，ROCグラフにより異なる閾値を分類子としたパフォーマンスを視覚的に表すことができた．また，ROC凸包を用いて，検索対象となるキーワードの出現頻度や検索漏れに関するエラーコスト比に応じた理想的なシステム閾値を効果的に決定することが可能になると考えられる．今後，ROC凸包による閾値の決定以外の方法も導入して，支援システムとして適切な提示関連キーワード数を決定するアルゴリズムの開発が必要である．

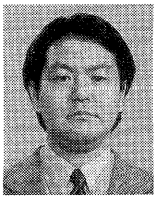
謝辞 本論文の一部は，文部省科学研究費重点領域における「分散発展型データベースシステム技術の研究(08244103)」での研究成果による．全文検索システムOpenTextl実行環境の提供をいただいた日商岩井インフォコムシステムズ(株)，日本サン・マイクロシステムズ(株)，伊藤忠テクノサイエンス(株)に感謝する．日ごろ御指導いただく南山大学経営学部情報管理学科 長谷川利治教授に深謝する．システム構築を支援して頂いた京都大学大型計算機センター 永平廣則氏に感謝する．最後に，本論文に対して貴重かつ有益な御指摘並びにコメントをいただいた査読委員の方々に感謝する．

参考文献

1) Barber, C., Dobkin, D. and Huhdanpaa, H.: The quickhull algorithm for convex hull, *Thechnical Report GCG53*, University of Minnesota (1993).
2) 川原 稔, 河野浩之, 長谷川利治: 文献データベース情報検索に対するデータマイニング技術の適用, 情報処理学会論文誌, Vol. 39, No. 4, pp. 878–887 (1998).
3) Kawahara, M., Kawano, H. and Hasegawa, T.: Implementation of Bibliographic Navigation System with Text Data Mining, *Proc. of XIII Int'l Conf. on Systems Science*, Poland (1998).
4) 川原 稔, 河野浩之: 文献情報検索支援システムにおける相関ルール選択基準, 情報処理学会研究報告, 98-DBS-116(1), pp. 135–142 (1998).
5) 河野浩之, 川原 稔, 長谷川利治: 文書データマイニングによる雑誌記事索引データベース検索支援, 情報学シンポジウム, pp. 121–128 (1998).
6) 河野浩之: データベースからの知識発見の現状と動向, 人工知能学会誌, Vol. 12, No. 4, pp. 497-504 (1997).
7) Kowalski, G.: *Information Retrieval Systems*, Kluwer Academic Publishers (1997).
8) Michalski, R. S., Bratko, I. and Kubat, M. (eds.), *Machine Learning and Data Mining, Methods and Applications*, John Wiley & Sons, Inc. (1998).
9) Parsaye, K., Chignell, M., Khoshafian, S. and Wong, H.: *Intelligent Databases*, John Wiley & Sons, Inc. (1992).
10) Provost, F. and Fawcett, T.: Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions, *Proc. of 3rd Int'l Conf. on Knowledge Discovery and Data Mining (KDD-97)*, pp. 43-48 (1997).
11) Salton, G.: Another look at automatic text-retrieval system, *Comm. ACM*, Vol. 29, pp. 648–656 (1987).
12) Srikant, R. and Agrawal, R.: Mining Generalized Association Rules, Dayal, U., Gray, P. M. D. and Nishio, S. (Eds.), *Proc.21st VLDB*, pp. 407–419, Zurich, Switzerland (1995).

(1998年 9月 20日受付)
(1998年 12月 27日採録)

(担当編集委員 藤原 譲)



川原 稔(正会員)

昭和63年3月早稲田大学理工学部電気工学科卒業. 平成2年3月京都大学大学院工学研究科応用システム科学専攻修士課程修了. 同年4月同大学大型計算機センター助手. 平成7年4月同大学院工学研究科応用システム科学専攻助手兼任. 平成10年4月同大学院情報学研究科システム科学専攻助手兼任. データベースシステム, データマイニングの研究に興味をもつ. 人工知能学会会員.

河野 浩之(正会員)

昭和60年3月京都大学工学部数理工学科卒業. 平成2年3月同大学院工学研究科数理工学専攻博士課程研究指導認定退学. 同年4月同大学工学部数理工学教室助手. 同時に同大学院工学研究科応用システム科学専攻助手兼任. 平成5年カナダ・サイモンフレーザー大学においてデータベースシステムの研究に従事. 平成8年4月京都大学大学院工学研究科応用システム科学専攻助手. 平成9年10月同大学院工学研究科応用システム科学専攻助教授. 平成10年4月同大学院情報学研究科システム科学専攻助教授. 工学博士. 情報伝送システム, データベースシステムの研究に興味をもつ. ACM, IEEE, AAAI, 電子情報通信学会, 人工知能学会各会員.