データベースシステム 116－18
（1998. 7. 8）

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*

川原 稔† 河野 浩之‡

†京都大学大型計算機センター ‡京都大学大学院情報学研究科

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，本稿では，ROC (Receiver Operating Characteristics) 空間を利用して，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．その結果，ROCグラフを用いることで，より検索精度の高い相関ルール導出を行える閾値の決定を行うことが可能となった．

キーワード：文献検索，データマイニング，相関ルール，ROC空間

Performance Evaluation of Bibliographic Navigation System with Association Rules in ROC Space

Minoru Kawahara† Hiroyuki Kawano‡

†Data Processing Center, Kyoto University
‡Department of Systems Science, Kyoto University

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 space, we can estimate appropriate threshold values to derive association rules for keywords.

Keywords: bibliographic search, data mining, association rule, ROC space

* 連絡先: 〒606-8501 京都市左京区吉田本町 京都大学大型計算機センター 川原稔
Tel: (075)753-7429, E-mail: kawahara@kudpc.kyoto-u.ac.jp

1 はじめに

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，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．このような困難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関わる研究が数多く行われている [2, 4, 5, 7, 9]．そこで，我々は，データマイニング手法 [3, 6] の一つである相関ルール（association rule）導出アルゴリズム [10] 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, 4]．そのシステムは，図 1 のように，導出されたルールに基づいた関連キーワードを検索ユーザに提示することで，検索に関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，複数属性の関係から導出に関わる各種閾値を動的に決定している [2]．アルゴリズムにより関連キーワードが導出されない場合は，最小サポート閾値 $Minsup$ および最小確信度閾値 $Minconf$ を緩和して，関連キーワードが導出されるようにしている．逆に，関連キーワードが多数導出された場合には，導出される関連キーワード数がシステムの設定値 $Maxkey$ 以下となるように，閾値を厳しくして関連キーワードの導出を抑制している．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，本稿では，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) 空間 [1, 8] を用いて，検索要求に対する相関ルール導出に関わる閾値を視覚的に評価する方法について議論する．

以下，2 章で，ROC 空間についての概略を述べる．3 章では，ROC 空間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基づいたパラメータを設定し，ROC 空間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．

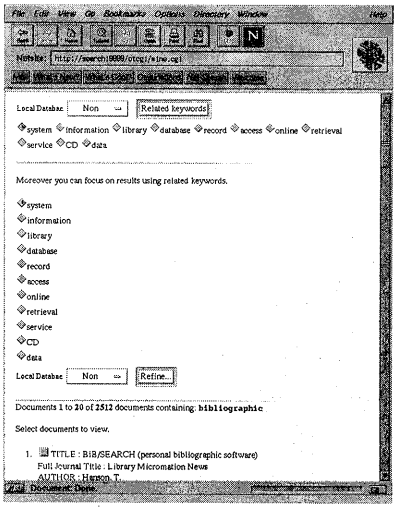

図 1: 文献情報検索支援システムの検索結果画面

2 ROC 空間

コスト分布やクラス分布を正確に把握することが困難な場合に，ROC 空間を用いて分類子のパフォーマンスを視覚化することで，分類子のパフォーマンスを比較することが可能になる [1, 8]．

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，事象 I が分類 c となる事後確率を $p(c \mid I)$ で表すと，正の事象 P が正 y と正しく分類される比率 TP は，

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

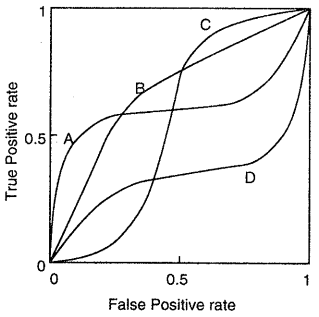

図 2: 4つの異なる分類子の ROC カーブ

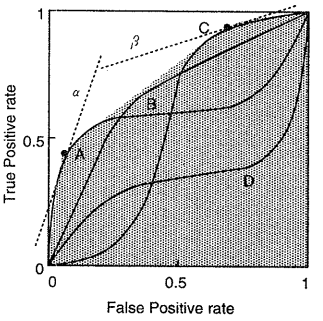

図 3: ROC 凸包における等パフォーマンス線

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，

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

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，たとえば，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 空間では，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” および “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 空間上の点 (FP, TP) に対する分類子のコストは，

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

により表されることになる．

ここで，ROC 空間における 2 点 (FP_1, TP_1) および (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，

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

となる．つまり， この等式は等パフォーマンス直線 (iso-performance line) の傾きを与えている．たとえば，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関わらず 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関わらず n と分類されてしまうコストが等しいときの傾きは 3 となり，10 倍コストがかかるときの傾きは 3/10 となる．これらは，図 3 でそれぞれ α および β のように表される．

3 文献情報検索支援システムへのROC空間の適用

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空間を適用できるよう，$\sum$ を集合の論理和演算子として以下を定義する．

[定義] k_i : i 番目の検索要求キーワード．
$\mathbf{K}_i$: k_i が被覆する文献の集合．
m : k_i の関連キーワード数．
$r_{i,j}$: k_i の j 番目の関連キーワード．
$(1 \leq j \leq m)$
$\mathbf{R}_{i,j}$: $r_{i,j}$ が被覆する文献の集合．
$|\,\mathbf{S}\,|$: 集合 $\mathbf{S}$ に含まれる文献数．

図 4 は，検索対象となる文献データ全ての集合 $\mathbf{U}$ に対するキーワード $\mathbf{K}_i$ および $\mathbf{R}_{i,j}$ による被覆状態を示している．

絞り込み検索においては，$\mathbf{K}_i$ を絞り込む事象が正となり，逆に拡大する事象 $\overline{\mathbf{K}_i}$ が負の事象となる．したがって，正の事象でありかつ正と分類される事象は $\mathbf{K}_i \cap \sum_{j=1}^{m} \mathbf{R}_{i,j}$ であり，負の事象でありかつ正と分類される事象は $\overline{\mathbf{K}_i} \cap \sum_{j=1}^{m} \mathbf{R}_{i,j}$ であるから，TP は，

$$TP = \frac{|\,\mathbf{K}_i \cap \sum_{j=1}^{m} \mathbf{R}_{i,j}\,|}{|\,\mathbf{K}_i\,|}$$

で表すことができ，FP は，

$$FP = \frac{|\,\overline{\mathbf{K}_i} \cap \sum_{j=1}^{m} \mathbf{R}_{i,j}\,|}{|\,\overline{\mathbf{K}_i}\,|}$$

で表すことができる．たとえば，図 4 で，正の事象でありかつ正と分類される事象が $\mathbf{K}_i \cap \mathbf{R}_{i,2}$ であり，負の事象であるが正と分類される事象が $\overline{\mathbf{K}_i} \cap \mathbf{R}_{i,2}$ である．

逆に，広がり検索においては，$\mathbf{K}_i$ を拡大する事象 $\overline{\mathbf{K}_i}$ が正となり，逆に絞り込む事象 $\mathbf{K}_i$ が負の事象となる．したがって，正の事象でありかつ正と分類される事象は $\overline{\mathbf{K}_i} \cap \sum_{j=1}^{m} \mathbf{R}_{i,j}$ であり，負の事象でありかつ正と分類される事象は $\mathbf{K}_i \cap \sum_{j=1}^{m} \mathbf{R}_{i,j}$ であるから，TP は，

$$TP = \frac{|\,\overline{\mathbf{K}_i} \cap \sum_{j=1}^{m} \mathbf{R}_{i,j}\,|}{|\,\overline{\mathbf{K}_i}\,|}$$

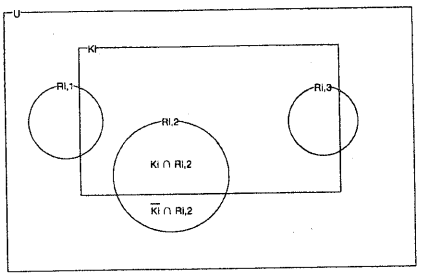

図 4: 文献情報空間における被覆関係

となり，FP は，

$$FP = \frac{|\,\mathbf{K}_i \cap \sum_{j=1}^{m} \mathbf{R}_{i,j}\,|}{|\,\mathbf{K}_i\,|}$$

となる．丁度，絞り込み検索の場合と TP と FP が逆に扱われることになる．

この TP と FP を基に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 [8] 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．なお，本稿では，数多い文献情報の検索結果から目的の文献を有効に絞り込む方法について考察するため，絞り込み検索の場合について論じる．

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．これらのデータを基にして，相関ルールを導出して関連キーワードによる絞り込み検索の支援を行っている．本章では，このデータを基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 $330,562$ となる．

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[1]，頻

[1] 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．

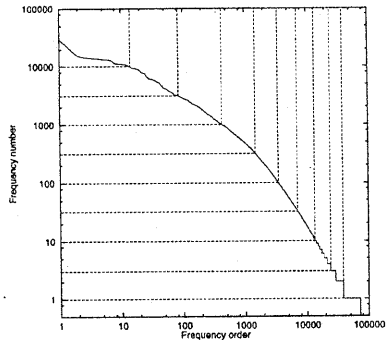

図 5: キーワード出現頻度

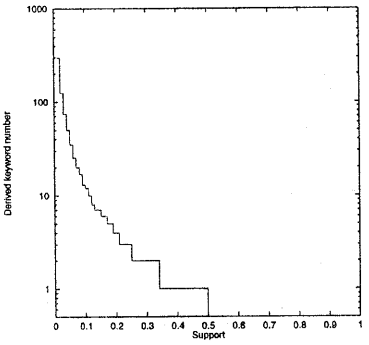

図 6: 平均導出キーワード数

表 1: 検索キーワードのカテゴリ

	出現回数	頻出順位	キーワード数
1	10001-	1 - 13	13
2	3163-	14 - 83	70
3	1001-	84 - 424	341
4	317-	425 - 1472	1048
5	101-	1473 - 3446	1974
6	33-	3447 - 7008	3562
7	11-	7009 - 13310	6302
8	4-	13311 - 24032	10722
9	2-	24033 - 38572	14540
10	1-	38573 - 73310	34738

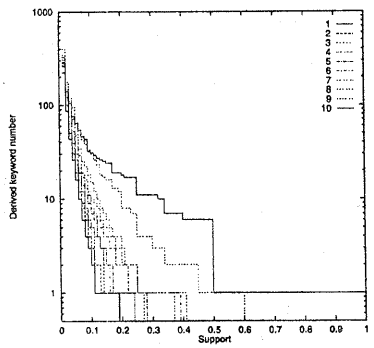

図 7: カテゴリごとの平均導出キーワード数

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．そこで，ROC 空間上にカーブが描かれるように，出現回数によりキーワードを表 1 のようなカテゴリにクラス分けして (FP,TP) を求める．ちょうど，各カテゴリは図 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．それぞれのカテゴリから各々 20 個のキーワードをサンプリングして，各カテゴリにおける (FP,TP) の平均を求めてプロットを行った．

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関連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，カテゴリごとの平均導出キーワード数は図 7 のようになった．

$Minsup$ を分類子とした ROC カーブを描くため，これを基にして，各カテゴリにおける導出キーワード数が 0 となる近傍のサポート値と，全てのカテゴリにおいてキーワードが導出されるサポート値について適宜刻みを取り，

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

について評価を行った．なお，このときの導出キーワード数は図 8 のようになった．

図 8 を見ると，同じ $Minsup$ の値を取っていても，出現頻度が低いキーワードからは，導出される関連キーワード数が多くなる傾向があることが分かり，そこから，ROC 空間においても，文献空間を被覆するキーワードが増えるため，TP 値が高くなることが予測される．そこで，各カテゴリにおける $Minsup$ に対する

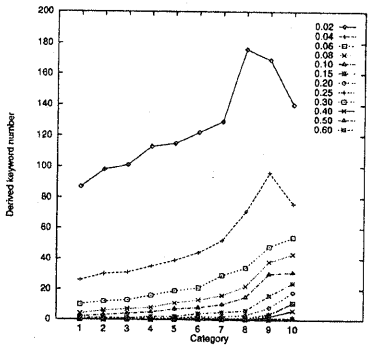

図 8: *Minsup* による平均導出キーワード数

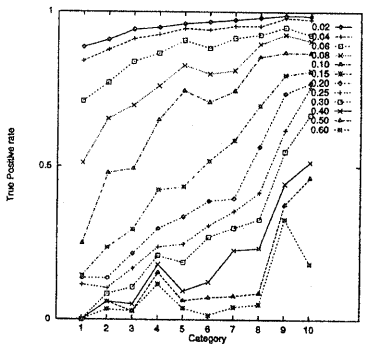

図 10: カテゴリに対する *TP* 値

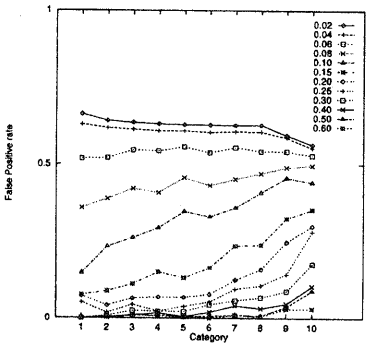

図 9: カテゴリに対する *FP* 値

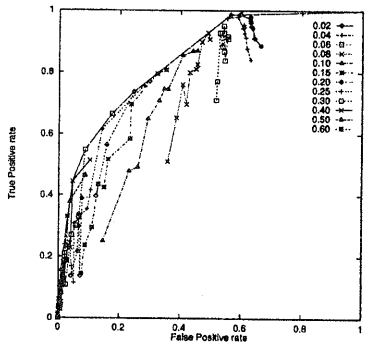

図 11: *Minsup* を分類子としたROCグラフ

FP および*TP* を調べると，それぞれ図 9 および図 10 のようになり，予測通り *TP* 値が高くはなっていることが確認できた．しかし，多くの関連キーワードを提示されたのでは，有効な絞り込み検索を行うことができないため，出現頻度が低くなるに連れて *Minsup* を厳しくして導出されるキーワード数を抑制する必要があることが分かる．

各 *Minsup* の値に対する各カテゴリの (FP, TP) を求めて，ROC 空間にプロットしたものが図 11 である．図 11 には，ROC グラフから ROC 凸包を用いて各領域における最適分類子を求め，等パフォーマンス線を描いている．図 11 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* の値を求めると，表 2 が得られる．

ここで，各カテゴリの $p(N)/p(P)$ の平均値を検索キーワードのヒット件数から求めると，表 3 の "$p(N)/p(P)$" のようになる．たとえば，絞り込み検索における被覆漏れが無いように，負の事象を正と判断するコストに対して正の事象を負と判断するコストを 10,000 倍，すなわち，$c(y,N)/c(n,P) = 10000$ とした場合のコスト比は表 3 の " コスト比" のようになり，最適のパフォーマンスを示す *Minsup* は "最適 *Minsup*" のようになる．なお，表 2 および表 3 で，"AllPos" はすべての関連語を導出することを示し，"AllNeg" は何も関連語を導出しないことを示している．ここから，先に述べた，出現頻度が低いキーワードに対しては，ROC 凸包では *Minsup* が厳しくなるため，導出されるキーワード数も抑制されることが分かる．

ROC 凸包により導出される関連キーワード

表 2: ROC 凸包による最適 $Minsup$

適応範囲	分類子
0.0000 – 0.0302	AllPos
0.0302 – 0.0416	0.02
0.0416 – 0.7875	0.02
0.7875 – 0.9921	0.20
0.9921 – 1.3338	0.30
1.3338 – 2.5398	0.30
2.5398 – 6.9658	0.40
6.9658 – 8.4813	0.60
8.4813 – 14.702	0.50
14.702 – 16.343	0.40
16.343 – 173.72	0.60
173.72 – Inf	AllNeg

表 3: 各カテゴリの $p(N)/p(P)$ とコスト比

カテゴリ	$p(N)/p(P)$	コスト比	最適 $Minsup$
1	25	0.0025	AllPos
2	75	0.0075	AllPos
3	207	0.0207	AllPos
4	640	0.0640	0.02
5	1952	0.1952	0.02
6	6254	0.6254	0.02
7	18206	1.8206	0.30
8	58365	5.8365	0.40
9	145997	14.5997	0.50
10	330561	33.0561	0.60

の傾向を調べるため，例として，$c(y,N)/c(n,P) = 10000$ とした場合の各種の値を求める．まず，求まった最適 $Minsup$ を用いて関連キーワードを導出した場合に，関連キーワードが導出されない割合を求めたものが図 12 であるが，非導出率はカテゴリにより大きく異なっており，図 13 および図 14 のようになっている．

以上の結果から，ROC 凸包で求めた最適 $Minsup$ による導出結果をまとめると，表 4 のようになった．

表 4 を見ると，キーワードの出現頻度に応じた $Minsup$ が与えられることが分かる．しかし，関連キーワードが大量に導出されるか，あるいは，ほとんど導出されない結果となってしまった．これは，図 8 に見られるような，大量の関連キーワードを導出する緩い $Minsup$ を ROC 空間に取り込んでいるため，TP が高い

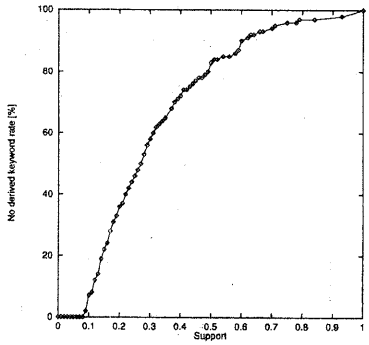

図 12: 関連キーワード非導出率

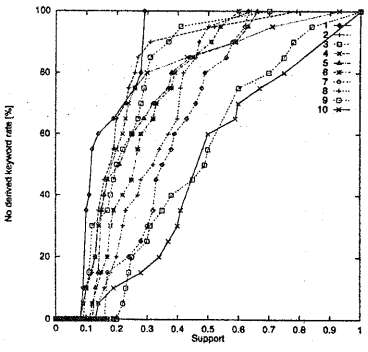

図 13: $Minsup$ に対するカテゴリごとの関連キーワード非導出率

値 (AllPos) に引きずられていることと，図 13 や図 14 に見られるような，関連キーワードが導出されない確率が高い $Minsup$ が FP を低い値 (AllNeg) に引きずられていることが原因と考えられる．このような分類子を有効に除外する方法についても考察していく必要があると考えられる．

5 結論

相関ルール導出における異なる閾値による ROC グラフを用いることにより，設定する閾値による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

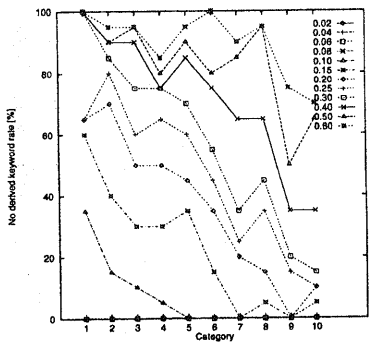

図 14: カテゴリに対する *Minsup* ごとの関連キーワード非導出率

表 4: ROC 凸包からの導出結果

カテゴリ	導出キーワード数	非導出率 [%]
1	261	0
2	281	0
3	290	0
4	113	0
5	115	0
6	122	0
7	1	30
8	0	65
9	1	55
10	1	70

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

本稿の一部は，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．また，日頃御指導頂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．全文検索システム OpenText 実行環境を提供いただいた日商岩井インフォコムシステムズ（株），伊藤忠テクノサイエンス（株），日本サン・マイクロシステムズ（株）に感謝する．最後に，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の永平廣則氏に感謝する．

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] 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.

[4] 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.

[5] Kowalski, G., "Information Retrieval Systems," Kluwer Academic Publishers, 1997.

[6] Michalski, R. S., Bratko, I. and Kubat, M., (eds.), "Machine Learning and Data Mining, Methods and Applications," John Wiley & Sons, 1998.

[7] Parsaye, K., Chignell, M., Khoshafian, S. and Wong, H., "Intelligent Databases," John Wiley & Sons, Inc., 1992.

[8] Provost, F. and Fawcett, T., "Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions," Proc.of KDD-97, pp.43-48, 1997.

[9] Salton, G., "Another look at automatic text-retrieval system," Comm. ACM, Vol.29, pp.648–656, 1987.

[10] 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.