テキストマイニングを用いた文書検索システム

○河野 浩之
川原 稔

Document Navigation System using Text Mining Algorithms

○ Hiroyuki KAWANO
Minoru KAWAHARA

大量に蓄積されつつある電子化データに対して，機械学習，データベース，統計などを基礎としたデータマイニング技術を応用した知識発見ツールが盛んに開発されている．また，我々は，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「問答」の構築を行っている．そこで，本稿では，ハイパーテキストである web 文書，INSPEC データベース，国会図書館雑誌記事索引データなどの大量の文書データに対してテキストマイニング技術を適用した実験結果について論じる．まず，実時間性のある検索支援を行うための効率的なルール導出戦略に関する議論を行う．次に，ヒューリスティックに与えられる閾値と，導出される相関ルールの関係について論じた上で，検索精度の優れた相関ルールを導出する閾値決定法に関して，ROC (Receiver Operating Characteristics) グラフを利用しながら述べる．

Many knowledge discovery tools have been developed using data mining, the integrating technologies of machine learging, database, statistics and others. We have been constructing "mondou" search systems based on extended association rules. In this paper, we discusse the experimental results of text mining applied for web hyper texts, INSPEC database, and magazines and articles index data in the National Diet Library. First of all, we express about the efficient strategies in order to derive association rules. Next, we discuss the relation between the threshold values and association rules, and we focus on the techniques of ROC (Reciever Operating Characteristics) 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

データマイニング (data mining)，もしくは，データベースからの知識発見 (KDD: Knowledge Discovery in Databases) は，機械学習，データベース，統計学などのデータ分析に関わる理論と技術を，大規模でノイズや例外を含む実データに対して適用する研究分野である [4]．そして，市場データや医療データなどの実データをはじめとして，大量に蓄積されつつある多種多様な文書データに対する適用も行われており，テキストマイニング (text mining) と呼ばれる分野を形成しつつある[1]．

例えば，ハイパーテキストなどの電子化文書データに対する決定木 (decision tree)，相関ルール (association rule)[5]，クラスタリング (clustering) などのアルゴリズムの適用，さらに，求められたルールの視覚化 (visualization) 技術を用いて，大量文書の特徴把握を支援する研究である．

なお，テキストマイニングを利用したシステム構築に関わる研究を深めることによって，図書・文献データベースを含む各種情報検索における問題を効果的に緩和し解消しうる可能性がある．

そこで，我々は，jp ドメインの web サーバ上にある 100 万ページ以上のハイパーテキスト文書に対して，データマイニングアルゴリズムの一つである相関ルールを拡張して検索支援に用いる「問答」システムを開発した．その後，「問答」を，INSPEC データベース，国会図書館雑誌記事索引データなどの大量の文書データに対しても適用する拡張を行ない，商用文書検索システム OpenText と連携させながら実験を行なっている．そして，現在，上述した文書データに対する検索支援システムのプロトタイプ実験を繰り返し，半構造データに対する検索支援システムの実証実

[1] シフトウェア (siftware) と総称されるツールは，機械学習，統計処理，回帰分析 (regression analysis)，帰納学習，ニューラルネット，対話型のデータ検索などを含む．"http://www.kdnuggets.com/siftware.html"

験を行うに至っている [1, 2, 3].

以下，第2章においてテキストマイニング研究の動向を簡単に紹介し，第3章で検索支援システム「問答」に関わる幾つかの構築技術を紹介する．

2 テキストマイニング

データマイニングは，生データを中心にした知識発見であり，単純明快な法則で記述し難い領域のデータを中心に扱う技術である．そして，より質の高い知識を求めるには，知識発見に関わる数多くのステップのバランス良い実行，また，収集された生データに対する前処理と，求められたルールに対する後処理が必要とされている．

そして，既存の文書データベースシステムが提供している問合せ処理機構では，増大する電子化文書に対して十分な処理能力を備えているとは言い難く，新たな技術を必要としている．そこで，既存の文書検索システムに実装されていないテキストマイニング技術に焦点を当てることによって，文書データの組織化，内容要約，意味抽出などの実用化が試みられている．また，文書データベースに対して対話的かつ視覚的な検索機能を追加するだけではなく，再現率 (recall) や適合率 (precision) と異なる基準を用いて検索性能の評価を進める必要がある．

なお，テキストマイニング研究を精力的に行っている Ronen Feldman らは，「データを意味付ける」ソフトウェアとして "Document Explorer" を開発するとともに，「キーワード分布，背景知識，データ構造，相関性，漸増的アルゴリズム，テキスト構造，特徴抽出，大量文書処理」などの問題からテキストマイニング研究へとアプローチしている．なお，Document Explorer は，個々の文章の特徴を抽出するだけではなく，文書データが記述されたタイムスタンプを利用した傾向変化，さらに，文書集合間の関連を明確にする機能提供，文脈グラフ（context graphs），傾向グラフ（trend graphs），カテゴリー関連マップ（category connection maps）などを提供するものである．

3 問答

本章では，実時間性の高い検索支援を行うための効率的なルール導出戦略に関する議論を行う．次に，ヒューリスティックに与えられる閾値と，導出される相関ルールの関係について論じた上で，検索精度の優れた相関ルールを導出する閾値決定法に関して，ROC (Receiver Operating Characteristics) グラフを利用しながら述べる．

3.1 ルール導出戦略

データマイニング技術を応用したルール導出における問題点として，その処理にかかわる計算コストが比較的大きくなることがある．この種の計算コストにかかわる問題に対して，データマイニング [4] の研究初期のころから，サンプリング手法等が考えられているが，サンプリング手法は必ずしもルールの正確さが保証されないなど問題も多い．そこで，ルール導出コスト削減の一つの選択肢として，文献データから抽出されたキーワードに対してあらかじめ検索時に必要となるルール導出処理を行う前処理手法が考えられる．この手法を実体化 (materialization)[1] と呼び，検索要求が予測されるキーワードに対して実体化を行い，その処理結果を高速アクセスが可能な領域 (プリセット領域) に配置して，実際に検索要求が発生したときにはプリセット領域から結果を引き出して用いる．

[定義]
- T: 検索システム観測期間
- n: 検索対象となるキーワード総数
- p: 相関ルール導出に用いる属性数
- k_i: 検索対象となるキーワード $(1 \leq i \leq n)$
- a_j: 相関ルール導出に用いる属性 $(1 \leq j \leq p)$
- $N_{j,i}$: 属性 a_j にキーワード k_i を含む文書の件数
- X_i: 観測期間 T にキーワード k_i が検索要求を受けた回数
- α: 補助記憶アクセスにかかわる処理を行うのにかかる単位時間コスト
- β: 主記憶アクセスにかかわる処理を行うのにかかる単位時間コスト
- m: 実体化を行うキーワード数
- U_i: 観測期間 T に更新されたデータに k_i が含まれる回数 (最初の実体化に必要なコスト含む)

N 個の文書をすべて主記憶にロードして相関ルール導出にかかる時間が $O(N \log N)$ であること，プリセット領域から結果を引き出すコストは実体化が行われていない場合にかかる時間コスト R_i に比べて微少であること，プリセット領域更新時の結果更新に要する時間コストは R_i に比べて微少であることを考慮すると，キーワード k_i の検索要求に対する時間コスト R_i，観測期間 T の総コスト C_T，検索実行時の実体化による時間コスト削減量と実体化の更新コ

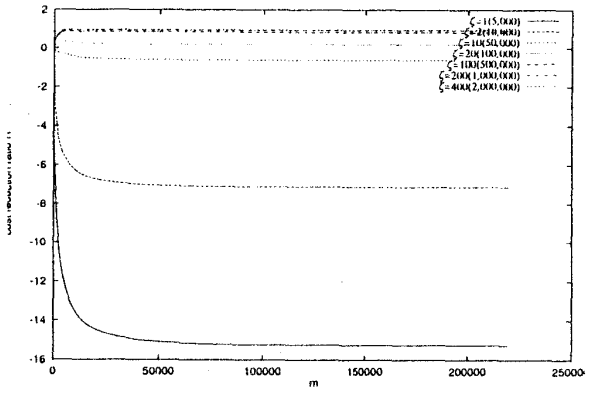

図 1: 均一分布の利得率 ($\beta/\alpha = 10^{-1}$)

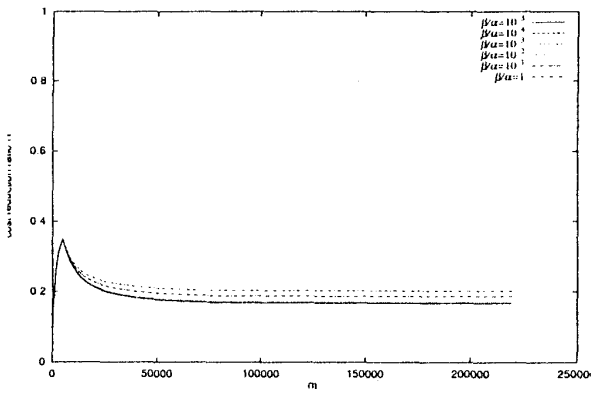

図 2: 均一分布の利得率 ($\zeta = 20$)

ストの差すなわち利得 G は,

$$R_i = \sum_{j=1}^{p}(\alpha N_{j,i} + \beta N_{j,i}\log N_{j,i}), \qquad C_T = \sum_{i=1}^{n} R_i X_i, \qquad G \simeq \sum_{i=1}^{m} R_i(X_i - U_i),$$

となる. コスト削減率を表す指標すなわち利得率 R は, 実体化による時間コスト削減量 G と実体化を採用しない場合の時間コスト C_T の比で次のように表せる.

$$R = \frac{G}{C_T} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{p}(N_{j,i} + \frac{\beta}{\alpha}N_{j,i}\log N_{j,i})(X_i - U_i)}{\sum_{i=1}^{n}\sum_{j=1}^{p}(N_{j,i} + \frac{\beta}{\alpha}N_{j,i}\log N_{j,i})X_i}.$$

評価対象として, 1987 年 1 月から 1996 年 12 月までの 10 年間に提供された 2,682,302 件の INSPEC データを用い, $N_{j,i}$ はこのデータを基に求めた. 更新データとして, 1997 年の INSPEC データを用いた. INSPEC データは月に 1 回の割合で追加提供されるので, 1997 年の 1 年間で更新が 12 回行われ, その間に提供された 330,562 件のデータを基にして初期投資コストと合わせて更新回数 U_i を求めた. したがって, 観測期間 T は 10 年, 検索対象となるキーワードはタイトル中に含まれるキーワードとし総数 n は 213,775 語となる. 相関ルール導出に用いる属性は, (タイトル, キーワード, アブストラクト, 著者) の 4 属性であるが, これにタイトルと著者の関連づけによるキーワード空間拡大処理を加え, $p = 5$ とする.

検索要求は出現頻度上位 5,000 語に対して均一に行われるものとして, $X_i = \zeta$ ($1 \le i \le 5000$, それ以外は 0) として, $\zeta = \{1, 2, 10, 20, 100, 200, 400\}$, $\beta/\alpha = \{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ とした場合の利得率を求めた. 図 1 は, $\beta/\alpha = 10^{-1}$ に値を固定して利得率を描いたものであり, 図 2 は, 100,000 件の検索要求に相当する $\zeta = 20$ に値を固定して利得率を描いたものであるが, β/α および ζ が他の値を取ったときも, それぞれ同様の傾向を示した.

図 1 を見ると $\zeta = 20$ (100,000 件相当) 以下の場合には, 利得率が落ち込んでいる. これは, 更新にかかわったキーワード数が延べ 575,019 語であり, 検索要求回数がこの値を下回ると利得率が落ち込むからであり, 実体化には更新キーワード数を上回る検索要求が必要であることがわかる. よって, 検索要求件数が 500,000 件以上, つまり $\zeta \ge 200$ の部分について議論を行う. また, 図 2 の利得率が落ち込み始めるまでの部分を見てわかるように, 利得率は β/α の影響をほとんど受けず, 実体化の効果にほとんど影響を及ぼさないことがわかる.

ここで, 利得率の特性を詳しく見るため, 図 1 の $m \le 5000$ の部分を拡大すると図 3 のようになる. 図 3 から, 全キーワード数 n の 0.01% 程度の 2,000 語の実体化を行うと, 全時間コストの 70%程度の削減ができ, 2,500 語の実体化で 80%程度を削減できることがわかる. また, それ以上の実体化を行っても, せいぜい 10%程度しか効率が上がらず, それ以上の実体化は不要ということになる.

3.2 導出ルールの評価

相関ルールを導出に用いる最小サポート閾値 $Minsup$ や最小確信度閾値 $Minconf$ 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が必要である. そこで, パフォーマンス空間の解析に有効な ROC 解析手法を用いて, 導出されたルールの評価を行い妥当な閾値を決定する [2].

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

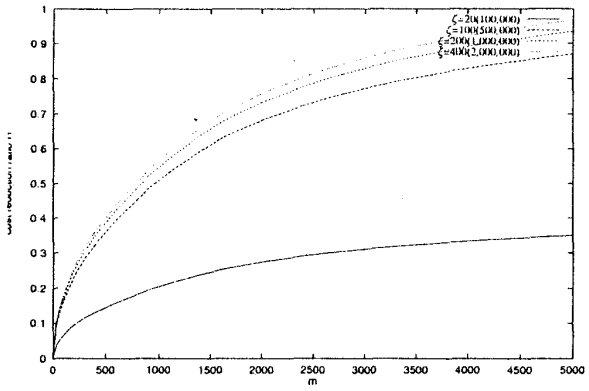

図 3: 均一分布の利得率 ($\beta/\alpha = 10^{-1}, m \leq 5000$)

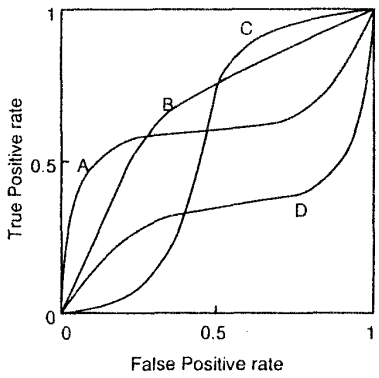

図 4: 4 つの分類子による ROC グラフ

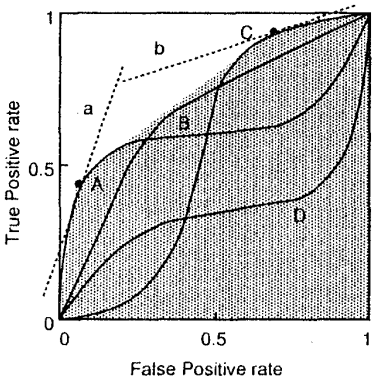

図 5: ROC 凸包における等パフォーマンス線

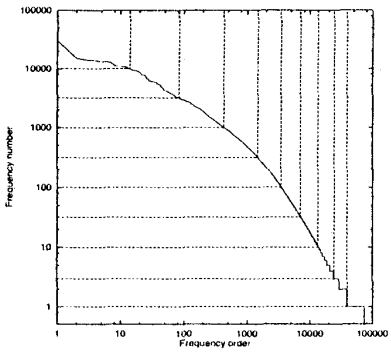

図 6: キーワード出現頻度

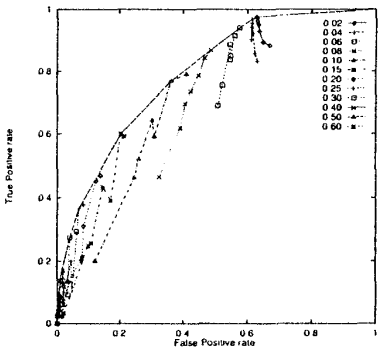

図 7: *Minsup* を分類子とした ROC グラフ

分類される比率 TP，および，負の事象 N が誤って正 y と分類される比率 FP は，次のように表すことができる．

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

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 4 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，グラフが左上端に近づくほど，すなわち，TP がより高く FP がより低くなるほど，分類子により事象が正確に分類されたことになる．

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し，正の事象の事前確率を $p(P)$ とすると，ROC グラフ上の点 (FP, TP) に対する分類子のコストは，

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

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，

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

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．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 5 の陰をつけた部分である．例えば，図 5 において，a で示される傾きの直線が接する 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．

表 1: ROC 凸包による最適 $Minsup$

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

表 2: $R_{error} = 145$ における $Minsup$

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

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　　$\mathbf{B}$: $\mathbf{G}$ により被覆される文献の集合

n: $\mathbf{G}$ の検索要求キーワード数　　m: $\mathbf{G}$ から導出されるキーワード数

k_i: $\mathbf{G}$ の i 番目の検索キーワード $(1 \leq i \leq n)$　　r_j: $\mathbf{G}$ から導出される j 番目のキーワード

$\mathbf{K}_i$: k_i が被覆する文献の集合　　$\mathbf{R}_j$: r_j が被覆する文献の集合

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 $\mathbf{B} = \bigcap_{i=1}^{n} \mathbf{K}_i$ 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 および FP は次のように表すことができる．

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

1997 年に INSPEC により配布された 330,562 件 ($|\,\mathbf{U}\,| = 330,562$) の文献データの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 6 のようになる．ROC グラフ上にカーブが描くため，出現回数に応じて図 6 の対数縦軸上でほぼ均等に分れるように，キーワードを出現頻度上位からカテゴリ 1〜10 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) の平均値をプロットする．なお，出現頻度が低いカテゴリ 8〜10 のキーワードでは有効な相関ルール導出ができないと考えられるので除外した．また，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，$Minsup = \{0.02, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2, 0.25, 0.30, 0.4, 0.5, 0.6\}$ と変化させ，$Minsup$ を分類子として各カテゴリの (FP, TP) を求めて ROC グラフにプロットしたものが図 7 である．ROC 凸包の境界線も同時に描いている．これに対して，ROC 凸包により最適 $Minsup$ を求めたものが表 1 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，$p(N)/p(P) = (|\,\mathbf{U}\,| - |\,\mathbf{B}\,|)/|\,\mathbf{B}\,|$ で表すことができる．ここで，エラーコスト比 $c(y,N)/c(n,P)$ を，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設定する．表 1 から，キーワード導出が行われない AllNeg となるのは，$p(N)/p(P) \cdot c(y,N)/c(n,P)$ の値が 227.06 以上のときであるから，

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

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．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 2 である．

表 2 の最小サポート閾値を用いると，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．したがって，本手法では非高頻出キーワードの導出がパフォーマンスを保証した上で可能となる．

4 むすび

これまで，文書データベースシステムの構築技術は，大量文書データの検索効率を考えた圧縮蓄積とインデックス生成，さらに，検索応答性能の向上に関わる研究などが行なわれた．そして，SGML, HTML, XMLなどの構造化文書を対象とした研究も活発に行なわれており，既に，商用システムOpenTextなどに実装されている．なお，インターネット上で提供されているサーチエンジンの性能などからも分かるように，数千万から数億の文書ファイルに対する実時間検索処理システムの構築が行われている．

一方，TRECや日本語文書に対するBMIR-J2(http://www.ulis.ac.jp/ ishikawa/bmir-j2/)のようなベンチマークテストを通じて情報検索に関わる研究も，その適用規模を拡大しつつある．そして，情報検索に関する国際会議であるSIGIRでも，インターネット上の検索エンジンに関する研究が取り上げられるなど，対象とする研究領域にも広がりを見せている[6].

従って，今後とも実データを扱いうる検索支援システムの高度化は強く要求されると考えられ，テキストマイニング技術もその一端を支えるものと考えられる．

謝辞

本稿の一部に，文部省科学研究費(10780259)の研究成果を含む．なお，全文検索システムOpenText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．

参考文献

[1] 川原稔, 河野浩之, "相関ルール実体化を行う文献情報検索支援システムの性能評価," 電子情報通信学会論文誌, Vol.J82-D-I, No.1, pp.165–173, 1999.

[2] 川原稔, 河野浩之, "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," 情報処理学会論文誌データベース, Vol.40, No.SIG3(TOD1), pp.105–113, 1999.

[3] 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.

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

[5] 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.

[6] Yamana, H., Tamura, K., Kawano, H., Kamei, S., Harada, M., Nishimura, H., Asai, I., Kusumoto, H., Shinoda, Y., and Muraoka, Y., "Experiments of Collecting WWW Information using Distributed WWW Robots," Proc. of SIGIR'98, Melbourne, Australia, pp.379-380, 1998.

河野 浩之 (京都大学大学院情報学研究科システム科学専攻)
Department of Systems Science, Kyoto University
川原 稔 (京都大学大型計算機センター)
Data Processing Center, Kyoto University