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

川原 稔[†] 河野浩之[††] 長谷川利治[††]

文献データベースに格納されているデータに関する領域知識や背景知識を持つ場合においても，一般に属性値の曖昧さのために効率良い検索は難しいことが多い．そこで，本稿では，データマイニングの分野で研究されている相関ルール導出アルゴリズムを拡張し，文献データベースの各属性の持つ属性値から検索領域に関するルールを導出する．また，属性間の関連を用いて相関ルールを求める対象空間を拡大し，導出されるルールの変化について考察する．さらに，特定領域に対して望ましい性質を持つルール集合を選択するために，複数のデータベースから相関ルールを求めてキーワード集合を選択する方法について述べる．なお，提案したアルゴリズムに基づく文献情報検索支援プロトタイプシステムを構築し，導出キーワード集合に関する性質などの評価も行う．

Data Mining Technologies for Bibliographic Navigation System

Minoru Kawahara,[†] Hiroyuki Kawano[††] and Toshiharu Hasegawa[††]

Without background and domain knowledge, it is generally difficult for naive users to retrieve appropriate results from bibliographic databases. The one of difficulties is due to the ambiguousness of keywords in search queries. In this paper, in order to provide knowledge of keywords space, we extend the one of data mining algorithms which discovers association rules from the bibliographic database, and we modify search queries using derived rules. Moreover, we have to refine derived rules, and we try to apply our proposed algorithm to different databases and derive different rule sets in the other view of different domains. We also develop a prototype of mining system for bibliographic navigation based on full text database, and evaluate the effectiveness of our proposed algorithms.

1. はじめに

図書・文献に関わるデータベースを用いた情報検索では，一般に検索領域に対する領域知識に加えて，検索システムに慣れ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，様々な情報検索システム構築に関わる研究が数多く行われてきている[10),11)]．

しかし，業務における計算機環境の急速な浸透は，文献や文献関連情報の電子化を加速しており，格納されるデータ量は著しい増加をみせている．よって，膨大なデータ処理が可能なアルゴリズムであることを保証したうえで，より良い文献情報検索システムを実現することが非常に重要な課題となっている．

また，こうした膨大なデータ処理が要求される研究は文献情報検索だけではなく，データマイニング（data mining）に関する研究で頻繁に論じられている[2),7)]．データマイニングは，データベースからの知識発見（KDD: Knowledge Discovery in Database）とも呼ばれており，様々な角度からさかんに研究されている．なお，膨大な文書データから実用性の高い興味あるルールを効率良く導くアルゴリズムは，文書データマイニング（text data mining）[3)]において研究されている[1),8),14)]．たとえば，文書データに対するアルゴリズムとして，自己組織化マップ（SOM: Self-Organizing Map）によるクラスタ化[9)]などが含まれる．

我々は，代表的なデータマイニングアルゴリズムである相関ルール（association rule）[13)]を拡張し，Webに関するテキストデータから導出されるルールを用いた検索支援を行うRCAAUシステムを開発している[4),8)]．また，提案したアルゴリズムの文献データベース情報検索への適用も試みている[5),6)]．

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

本稿では，キーワードやアブストラクトなどを含む文献データベースの諸属性に焦点をあて，効果的な検索支援を可能とする文献データベースを構成するアルゴリズムについて考察する．特に，単一属性から導出される相関ルールだけでなく，複数属性の関連を利用したルール導出によって，より好ましい検索を実現するための検索式改善アルゴリズムを述べる．加えて，文献データベースから多くの相関ルールが導出される場合に，より望ましいルールを選択する方法について考察する．たとえば，ユーザの嗜好の偏りを取り込むために，電子ニュースや電子メールなどから成る異種データベースをルール導出に援用することが考えられる．また，全文検索データベースに基づいた従来型の検索システムを基盤に，提案したアルゴリズムの実装を行い，その性能評価を行う．

以下，2章では，文献情報検索の現状と検索の困難さに関する簡単な議論を行う．3章では，相関ルール導出アルゴリズムにおけるアイテムを，キーワードの出現頻度を考慮するように拡張した重み付き相関ルール導出アルゴリズムについて述べる．4章および5章では，文献情報検索ユーザが有効な検索を遂行するうえで必要となるデータマイニングアルゴリズムの提案を行う．4章では，文献データベースにおける複数属性を利用することによりルール導出領域の拡大を試み，より良い関連ルール抽出の可能性を示す．5章では，検索ユーザにとって自然な手法でルール選択を実行するために，領域の異なるデータベースからのルール導出をあわせて用いるアルゴリズムについて述べる．さらに，6章では，4章および5章で提案したアルゴリズムを実装し，実装システムに関する性能等の評価について述べる．また，最後の7章において，結論と将来の課題について述べる．

2. 文献情報検索システムの問題点

文献情報データベースには，基本的に自由に作成された文書に関わるデータが格納されるため，通常のデータベースと異なって属性値の値域がまったく制限されない．さらに，データベース編纂者の分類方法によって属性や属性値も異なるうえに，著者や出版社により属性値の与え方も異なるため，より曖昧なデータが格納されがちである．そのため，文献情報検索ユーザは，検索において用いる属性や属性値を把握するのが困難となり，目的のデータを得るのが難しくなる[10]．

そこで，より優れた文献情報検索システム構築のために，索引付けやキーワード付与などを行うシステムが存在するが，組織や人に作業を頼っているため，索引付けなどの方法を完全に統制することは難しく，抽出データのゆらぎは避けられない[10]．また，文書ベクトル空間を用いた検索[11),12)]は，一般に文書ベクトル作成の手間が大きいうえに高い計算量を必要とするアルゴリズムが多く，大規模なデータベースに対する適用は現実的でないと考えられる．

さらに，検索対象となる電子化情報は増大する一途であり，また，近年のネットワーク環境の浸透は，文書検索に関わる背景知識や領域知識の不足した検索ユーザであっても，直接文献情報検索システムを操作する状況の急激な増加を招いている．そのため，検索に関わる知識の幅を広げる手法が重要であり，優れた分類（taxonomy），シソーラス（thesaurus），さらに概念木（conceptual tree）などの提供が要請される．しかし，組織や著者により単語の意味の位置付けが異なるため，単純にキーワードを検索式において展開することは，異なる観点のキーワードを混在させる可能性が生じる．また，データ量の増加と情報の複雑化の中で，分類・主題分析・キーワード統一などが難しくなっていることからも，大量に蓄積された文献情報からノイズや検索漏れを防ぐ検索支援システムの必要性は非常に高い．そこで，本稿では，INSPECデータを用いた図1の文献データベースに対して，各種文書検索技術を併用することを考慮した，より望ましい検索支援システムの設計を試みる．

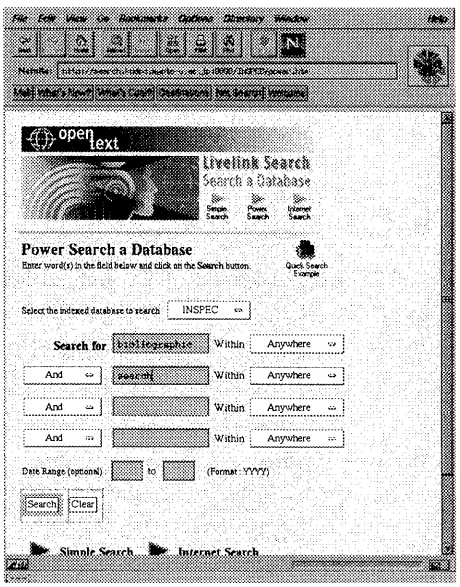

図1 全文検索システムを用いた文献情報検索システムの例
Fig. 1 Example of a full-text bibliographic retrieval system.

3. 重み付き相関ルール導出アルゴリズム

本稿では，多くの研究がなされている相関ルール導出アルゴリズム[13]を拡張し，検索に用いるキーワードを導出するアルゴリズムとして用いる．

まず，相関ルール導出アルゴリズムのキーワード集合のサポート値を，検索対象となる全体のタプル数に対するキーワード集合を含むタプル数の割合とする．次に，与えられたキーワード集合 $\mathbf{K_g}$ と，導出されるキーワード集合 $\mathbf{K_d}$ の要素によってデータ全体に対して検索を実行し，十分な検索需要が生じると判断される閾値 $Minsup$ 以上のサポート値を持つ，大きいキーワード集合を生成し，相関があるとされるキーワードを $\mathbf{K_d}$ へと加える．さらに，大きいキーワード集合から，$\mathbf{K_g}$ の検索要求がある場合に，$\mathbf{K_d}$ を用いた検索要求が同時に閾値 $Minconf$ 以上に生じると考えられるルールが，確信度の高いルールとして生成される．

ここで，検索に関係するすべてのタプル集合 $\mathcal{T}$ から，キーワード集合 $\mathbf{K} = \{k_j \mid j \in J\}$ を含むタプル $T_i (i \in I)$ を選択する．なお，キーワード集合 $\mathbf{K}$ は，検索に関係するすべてのキーワード集合 $\mathcal{K}$ の任意の要素のすべての組合せとなる．T_i における k_j の重みが w_{ij} であるとき，キーワード集合 $\mathbf{K}$ のサポート値 $sup(K)$ は次式によって求めることとする．

$$sup(K) = \frac{N(K)}{N_0},$$

$$N_0 = \sum_{\mathcal{T}} \max_{\mathcal{K}} w_{ij}, N(K) = \sum_{i \in I} \min_{j \in J} w_{ij}.$$

キーワード k_j で検索を実行したとき，あるタプル T_i において w_{ij} 個の重複がある場合，w_{ij} の重みがあるとし，キーワードと重みの組を $a_{ij} = (k_j, w_{ij})$ とする．

4. 複数属性からのルール導出

本稿で構築を試みる文献データベース情報検索支援システムは，検索対象となる属性に対しての相関ルール導出を行うだけではなく，各属性の位置付けを明確にし，それぞれに応じた相関ルール導出を行い利用する[6]．また，単独の属性からの相関ルール導出だけでなく，属性間の関係を用いた処理も行う[5]．

4.1 文献データベースにおける属性の特徴

キーワード空間を適切に拡大し異なる複数の属性の利用を考えるため，文献データベースにおける属性の特徴を整理する．属性は**表 1** のように 3 つのタイプに分類され，このうち相関ルール導出の対象として，特徴抽出属性と内容記述属性を用いる．

表 1 文献データベースにおける属性の分類
Table 1 Categorization of attributes in bibliographic database.

分類	内容
特徴抽出	文献の特徴を抽出した単語により構成される属性である．著者や出版社の他，データベース編纂者による値を与えられることがある． 例：タイトル，キーワード
内容記述	文献の内容に関して述べた単語により構成される属性である．文献の内容に限らず内容に関係する文献などについて値を与えられることがある． 例：アブストラクト，目次，索引
付加情報	文献の出版に付随する情報として与えられる属性である．直接的な内容に関しての情報は少ないが，文献に関する背景等についての情報が得られる． 例：著者，出版社，会議名

表 2 属性のキーワード空間
Table 2 Keyword space of the attribute categories.

分類	キーワード空間	INSPEC の平均単語数	
特徴抽出	小	タイトル	11
		キーワード	30
内容記述	小～大	アブストラクト	183
付加情報	小	著者	3

特徴抽出属性には，文献に対する高い相関性を持つタイトルやキーワードなどが含まれているが，属性値に含まれる単語数が少ないため，相関ルール導出アルゴリズムにより相関性の高いキーワードは求めにくい．そこで，関連する属性を利用してキーワード空間を拡大することを考える必要がある[5]．また，相関ルールによる関連キーワード導出では，“of”，“the”，“and”などの無意味語が多く導出されがちであるため，その除去も考慮しておく必要がある．そこで，本稿では，辞書を用いて無意味語の除去を行った後に，相関ルール導出アルゴリズムを適用することとする．

4.2 関連属性を用いたキーワード導出空間の拡大

表 2 に示すように，特徴抽出属性や付加情報属性は，一般的に含まれる単語数が少ないため，相関性の高いキーワードを数多く求めることは難しい．

そこで，**図 2** のようにキーワード空間を拡大することが考えられる．たとえば，属性 A_u を“タイトル”，属性 A_v を“著者”とすると，同一著者の著作物のタイトル集合からキーワード空間を構成することとなる．通常，文献情報検索ユーザは同一著者の著作物に関心を示すことが多いため，このような関連付けは妥当と考えられる．関連属性を用いたキーワード空間拡大ア

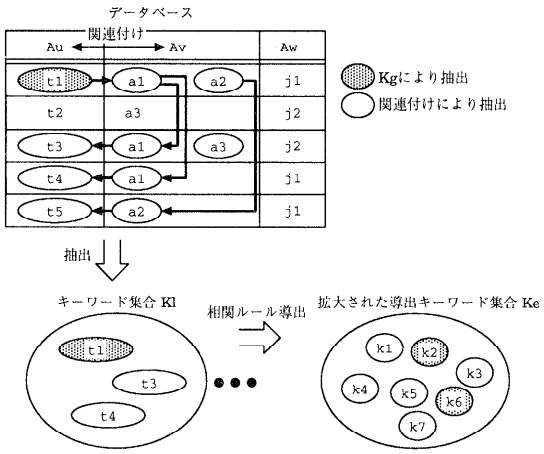

図 2 属性間の関連を用いたキーワード空間の拡大
Fig. 2 Extension of a keyword space using the relation between attributes.

ルゴリズムを示す.

アルゴリズム 1

入力： 初期検索式におけるキーワード集合 $\mathbf{K_g}$
特徴抽出属性 A_c
特徴抽出属性 A_u およびその関連属性 A_v

出力： 拡大された関連キーワード集合 $\mathbf{K_e}$

手順：

(1) 指定された属性 A_u が，与えられたキーワード集合 $\mathbf{K_g}$ に合致するタプル集合 $\mathbf{T_g}$ を抽出する.

(2) $\mathbf{T_g}$ から，A_u に関連付けされた属性 A_v によりキーワードを抽出してキーワード集合 $\mathbf{K_r} = \{k_l \mid l \in L\}$ を生成する.

(3) $\mathbf{K_r}$ の各キーワード k_l に対して，k_l を A_v に含むタプルの属性 A_u からキーワードを抽出して，キーワード集合 $\mathbf{K_l}$ を生成する.

(4) 集合 $\mathbf{K'_r} = \{\mathbf{K_l} \mid l \in L\}$ に相関ルール導出アルゴリズムを適用して，拡大された関連キーワード集合 $\mathbf{K_e}$ を導出する. □

4.3 複数属性を用いた検索式生成アルゴリズム

著者あるいはデータベース編纂者により与えられた特徴抽出属性は，高い相関性を示すキーワード集合 $\mathbf{K_c}$ を与える.

次に，アルゴリズム 1 を用いて，キーワード集合 $\mathbf{K_e}$ を導出する．特徴抽出属性を基にして得られるキーワード集合 $\mathbf{K_e}$ に含まれるキーワードは，属性間の関係を考慮しているため重要度が高い.

しかし，なお単語数の限られた属性値から得られるキーワード空間は狭い．そこで，内容記述属性から相関ルールを導出し，キーワード集合 $\mathbf{K_d}$ を求めてキーワード空間を拡大する．これにより得られるキーワード集合 $\mathbf{K_d}$ は，その文献の主題に関する以外の関連情報などの記述も含まれることがあり，より一般性が高いルールとなると考えられる．以上の特徴を考慮して，関連キーワードを導出するアルゴリズムを提案する.

アルゴリズム 2

入力： 初期検索式におけるキーワード集合 $\mathbf{K_g}$
特徴抽出属性 A_c
特徴抽出属性 A_u およびその関連属性 A_v
内容記述属性 A_d

出力： 関連キーワード集合 $\mathbf{K_o}$

手順：

(1) A_c における導出ルールからキーワード集合 $\mathbf{K_c}$ を得る.

(2) A_u と A_v の関連付けから，アルゴリズム 1 を適用してキーワード集合 $\mathbf{K_c}$ を得る.

(3) A_d における導出ルールからキーワード集合 $\mathbf{K_d}$ を得る.

(4) $\mathbf{K_e}$ と $\mathbf{K_d}$ の空でない共通部分集合 $\mathbf{K_i}$ が導出されるまで，最小サポート閾値 $Minsup$ および最小確信度閾値 $Minconf$ が緩和限界を切っていなければ，それらを緩和してステップ (1) に戻る.

(5) 論理和 $\mathbf{K_o} = \mathbf{K_c} \cup \mathbf{K_i}$ をとり，関連キーワード集合 $\mathbf{K_o}$ を出力する. □

なお，ステップ 4で，$Minsup$ と $Minconf$ の閾値はルール導出に応じて適宜緩和する．これは，関連キーワード導出では，多くのキーワードが提示されると有効なルールが埋没する恐れがあり，強い絞り込みは有効な情報まで排除する可能性があるからである．このように，一般ルール的な相関ルールから得られるキーワード集合において，絞り込みを緩和することにより，有効な情報をなるべく失うことなく，多くの関連キーワードの中から適切な関連キーワードを用いた検索式の生成の可能性が高まると考えられる.

5. 領域の異なるデータベースからのルール導出を用いた検索支援

文献データベースに与える検索キーワードとして，領域知識に含まれるか，あるいは，領域知識に近いキーワードを与えた場合，アルゴリズム 2 により関連するキーワードを得ることができる．しかし，与えたキーワードが一般性が高かったり，検索対象領域において特殊な単語であっても他の領域でも特殊な単語であるような場合，それぞれの領域において関連語が得られてしまうために，求めたいキーワードがそれらの中に

埋没してしまうことがある．たとえば，“system” や “analysis” という単語は様々な領域において用いられており，INSPEC データベースの約 200 万件の文献のタイトル中に，“system” は約 11 万件，“analysis” は約 5 万件使用されており，このような単語のみによる検索は不可能である．また，これらの単語から相関ルールを求めても，サポート値が全体に非常に小さくなるため，有効な関連語を得ることはできない．

そのような一般性が高い単語は，全体として見ると無意味語に近いものであるが，各領域においては意味を持つ単語も多い．しかしながら，一般性が高い単語しか指定できない検索ユーザが多い場合には，より望ましい領域を容易に指定する手法も必要である．そこで，特定領域に偏ったキーワード空間を構成して，そのキーワード空間から関連語を導出することにより，検索キーワードの改善を提示することができれば有効であると考えられる．

本稿では，ある種のデータベースが特定領域に偏ったキーワード空間を持つ[5]ことを仮定して，ユーザの要求する検索領域における関連キーワードを導出する方法を考える．たとえば，電子ニュースにおけるニュースグループや電子メールを用いたメーリングリストなどでは，特定の興味の対象となる事項に関する意見交換などが行われており，それらのキーワード空間は対象となる事項に偏っている．このように，文献データベースと独立に構築されたデータベースを，キーワード空間に取り込むことによって，検索ユーザは検索対象領域に関するデータベースを指定するだけで，検索領域を限ることができるようになる．

なお，対象領域として指定されるデータベースから導出されるキーワード集合 $\mathbf{K_h}$ は，検索ユーザの関心のある領域知識や背景知識を強く反映することが多いが，そのデータベースには含まれていないキーワードも多くなることにも注意を払う必要がある．以上を考慮して，領域の異なるデータベースを援用するアルゴリズムを提案する．

アルゴリズム 3

入力： 初期検索式におけるキーワード集合 $\mathbf{K_g}$
特徴抽出属性 A_c
特徴抽出属性 A_u およびその関連属性 A_v
内容記述属性 A_d
異種データベース D_h
最大導出キーワード数 $Maxkey$

出力： 関連キーワード集合 $\mathbf{K_o}$

手順：

(1) アルゴリズム 2 の (1) から (4) を実行する．

(2) 最終的に決定された $Minsup$ および $Minconf$ を用いて，D_h における導出ルールからキーワード集合 $\mathbf{K_h}$ を得る．

(3) 論理和 $\mathbf{K_o} = \mathbf{K_c} \cup \mathbf{K_h} \cup \mathbf{K_i}$ をとり，関連キーワード数 $|\mathbf{K_o}|$ が $Maxkey$ を超えていたら，緩和限界値を引き上げてステップ (1) に戻る．

(4) $\mathbf{K_o}$ を出力する． □

6. 検索支援システムの構成と評価

本章では，文献データベースとして INSPEC データベースを用いた情報検索支援システムの構成とアルゴリズムの評価を述べる．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代表的な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．本実験システムには，INSPEC より 1990 年 1 月から 1997 年 5 月の間に配布された 2,085,629 件の文献データを格納している．

6.1 システム構成

本システムは**図 3** のような構成であり，INSPEC データに対する通常の全文検索は，全文検索システム OpenText によって実現している．INSPEC データベースに対する文献情報検索は，Web ブラウザから OpenText への HTTP インタフェースである Livelink Search を介して OpenText に対して問合せを発行することで行う．Web ブラウザと Livelink Search を中継する CGI において，OpenText に対して問合せを発行して検索結果を取得すると同時に，相関ルール導出アルゴリズムを組み込んだ相関ルール導出エンジンからキーワード集合を取得し，両者を組み合わせて HTML の形式でブラウザに渡して検索式の改善方法を提示する．

文献データベースからの相関ルール導出用には，あらかじめ**表 3** に示した 4 つの属性に対して属性値のキーワード解析を行い，各文献におけるキーワードと重みの組の情報を持つデータベースを構築した．このうち，特徴抽出属性 “タイトル” のキーワード空間をアルゴリズム 1 を用いて拡大するため，“タイトル” に関連付ける属性として付加情報属性 “著者” を選択した．つまり，アルゴリズム 1 において，“タイトル” は A_u に対応し，“著者” は関連属性 A_v に対応する．これらにアルゴリズム 1 を適用してキーワード集合 $\mathbf{K_e}$ を求める．

また，領域の異なるデータベースとして，実験的に

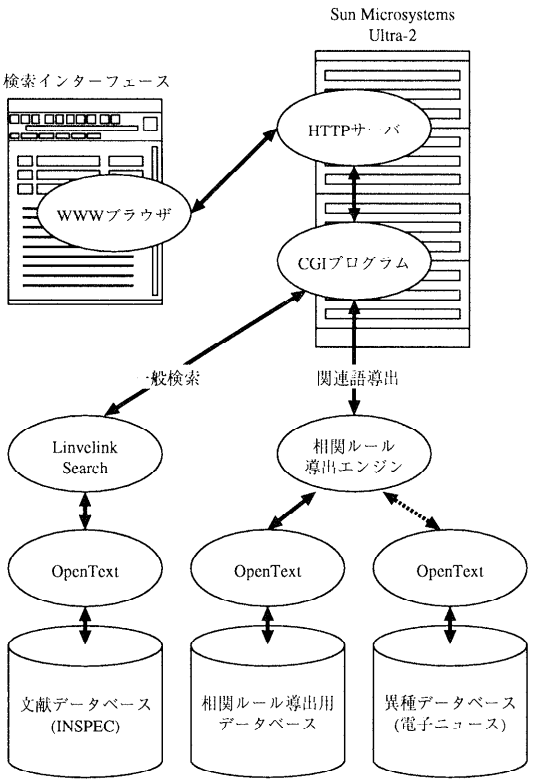

図3 本システムの構成
Fig. 3 The structure of bibliographic navigator.

表3 属性とキーワード集合の対応
Table 3 Correspondence attribute to keyword set.

属性	タグ	キーワード集合
タイトル	Title	$\mathbf{Kc_1}$
キーワード	Keyword	$\mathbf{Kc_2}$
著者（関連）	Title ⇒ Author ⇒ Title	$\mathbf{K_e}$
アブストラクト	Abstract	$\mathbf{K_d}$

表4 本システムで使用する異種データベース
Table 4 NetNews groups as the different resources.

分野	ニュースグループ	データ件数
計算機・制御・情報工学	comp.ai*	580
電子・電気工学	sci.electronics*	1,618
物理学	sci.physics*	1,519

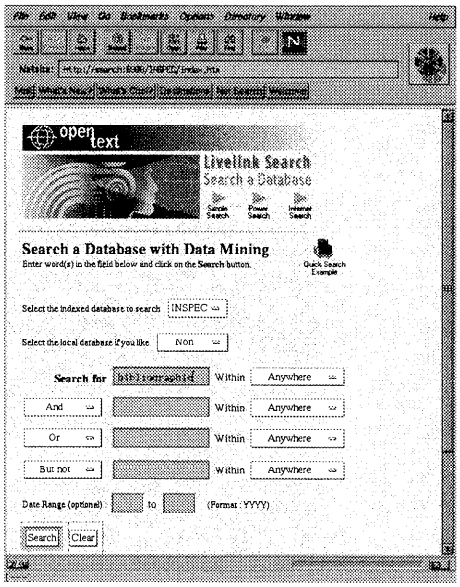

図4 本システムでの検索画面
Fig. 4 The query window of our bibliographic navigator.

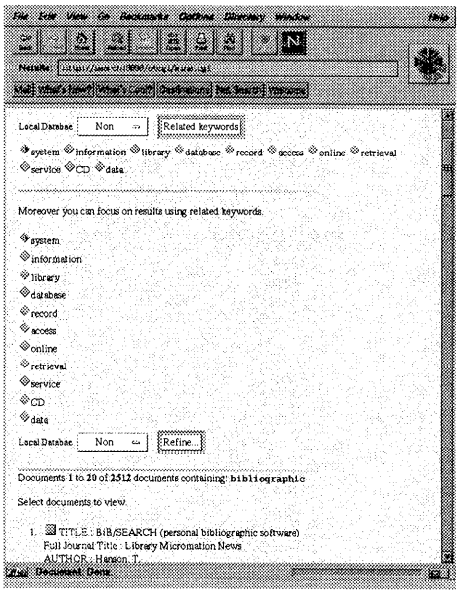

図5 本システムでの検索結果画面
Fig. 5 The result window of our system.

表4 に示すニュースグループを選択し，同様にキーワード解析を行い，各記事におけるキーワードと重みの組の情報を持つデータベースを構築した．検索時には，これらのデータベースを用いて相関ルールを導出する．なお，異種データベースの利用は検索ユーザにより選択可能とし，利用しない選択肢もある．すなわち，検索ユーザが異種データベースを選択しなければアルゴリズム2が，選択すればアルゴリズム3が実行される．

Webブラウザにおける検索画面は**図4**であり，検索結果として**図5**のような画面が表示される．検索結果画面には，検索結果がブラウザ上に提示されて，導出された関連語を用いて元の検索式に対して，絞り

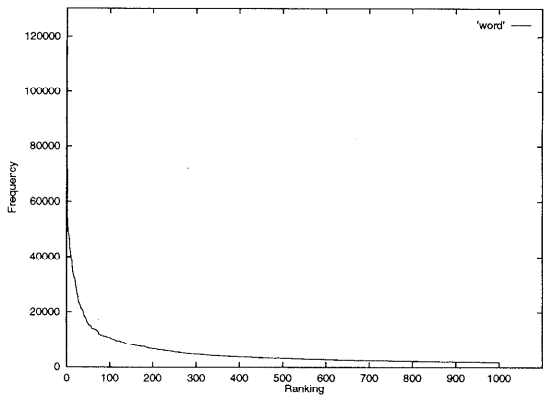

図6 最頻出1,000語の出現回数
Fig. 6 The frequency of the commonly used 1,000 words.

込み，または，関連語への移行の可能性を与え，選択ボタンにより自動的に検索式の改善が行われるようになっている．したがって，検索ユーザは，マウスのポインタ操作だけで改善方法を指定することができ，改善された検索式による検索を簡単に行うことができる．また，キーワード空間の性質の異なるデータベースを，画面上の選択ボタンにより選択することができるようにしてあり，相関ルールを導出するためのキーワード空間に偏りを与えることができる．

6.2 複数属性からのルール導出に対する評価

本システムで導出される関連語の傾向を調べるため，INSPECデータベースのタイトル部に使用される単語のうち，最頻出の1,000語について関連語の導出を行った．最頻出の1,000語に対して，出現順位と出現回数の関係は**図6**のようになっており，最頻出の100単語程度の出現回数が特に多いことが分かる．これらの単語から，それぞれ導出される関連キーワード数に対して，出現回数の多いものから順に100語ごとの平均を求めたものが**表5**である．表5を見ると，最頻出の100語の平均では，$\mathbf{K_e}$ と $\mathbf{K_d}$ の共通部分 $\mathbf{K_i}$ が得られていないという結果となっている．また，出現回数が多くなるほど，最終的に得られるキーワード $\mathbf{K_o}$ の数が少なくなる傾向がある．出現回数が多いにもかかわらず関連キーワード数が少ないということは，それらの単語が平均的に各領域のキーワード空間に分散して出現するため，サポート値が下がるためと考えられる．つまり，出現回数の多い単語は，特定の領域に偏ったキーワード空間を持っていないといえ，検索対象とするデータベース全体を被覆するキーワード空間を持つ一般性が高い単語といえる．

表5の結果から最頻出の100単語を除いた結果が**表6**のようになり，表6の多くの値が0となっていることから，最頻出100単語以外はほとんど導出されていなかったことが分かる．以上，アルゴリズム2を用いた場合，一般性が高い単語からは一般性が高い単語が関連語として導出されることが確認された．

表5 相関ルールより導出された関連キーワード数の平均
Table 5 Average number of the derived keywords.

頻出順位	$\mathbf{K_{c_1}}$	$\mathbf{K_{c_2}}$	$\mathbf{K_e}$	$\mathbf{K_d}$	$\mathbf{K_i}$	$\mathbf{K_o}$
1-100	1	2	28	2	0	3
101-200	2	3	38	3	1	4
201-300	2	3	35	4	1	5
301-400	2	3	44	4	2	5
401-500	2	4	49	4	2	5
501-600	2	4	58	5	2	6
601-700	3	5	44	6	2	7
701-800	2	4	41	4	2	5
801-900	2	4	46	4	2	6
901-1000	2	4	39	4	2	6
1-1000	2	4	42	4	2	5

表6 最頻出100語を除いた関連キーワード数の平均
Table 6 Average number of the derived keywords except for commonly used 100 words.

頻出順位	$\mathbf{K'_{c_1}}$	$\mathbf{K'_{c_2}}$	$\mathbf{K'_e}$	$\mathbf{K'_d}$	$\mathbf{K'_i}$	$\mathbf{K'_o}$
1-100	0	0	21	0	0	0
101-200	0	0	28	0	0	0
201-300	0	0	25	0	0	0
301-400	0	0	31	1	0	1
401-500	0	0	35	1	0	1
501-600	0	0	42	1	0	1
601-700	1	1	31	1	0	1
701-800	0	0	29	1	0	1
801-900	0	0	32	1	0	1
901-1000	0	0	27	1	0	1
1-1000	0	0	30	1	0	1

6.3 領域の異なるデータベース援用に対する評価

データベースA (comp.ai*) を援用してアルゴリズム3を最頻出の1,000単語に対して適用して，導出されたキーワード数の100語ごとの平均を求めたものが**表7**である．また，表7の結果に対して，最頻出の100単語を除いた結果が**表8**であり，表6に対応する．なお，データベースB (sci.electronics*) およびC (sci.physics*) を援用した場合も，表7および表8と同様の傾向を示す結果が得られている．

表5と表7の $\mathbf{K_e}$ および $\mathbf{K_d}$ の値を比べると，表7では表5の値に対して導出が抑制されている．また，表5の $\mathbf{K_o}$ に対する表6の $\mathbf{K'_o}$ の値と表7の $\mathbf{K_o}$ に対する表8の $\mathbf{K'_o}$ の値を比べると，導出されるキーワード集合における最頻出100単語が占める割合 $(1-\mathbf{K'_o}/\mathbf{K_o})$ が80%程度であったのに対して，領域の異なるデータベースの援用により40%以下におさえられている．これは，偏りのあるキーワード空間を持

表7 "comp.ai*" データベース援用時の関連キーワード数の平均
Table 7 Average number of the derived keywords using the "comp.ai*" database.

頻出順位	K_{c_1}	K_{c_2}	K_e	K_d	K_i	K_h	K_o
1-100	0	1	14	0	0	3	5
101-200	1	1	19	1	0	4	6
201-300	0	1	16	1	0	4	6
301-400	1	1	16	1	0	4	6
401-500	1	2	23	2	1	4	7
501-600	1	2	34	2	1	4	7
601-700	1	2	24	3	1	4	7
701-800	1	2	22	2	1	3	6
801-900	1	2	24	2	1	4	7
901-1000	1	2	21	2	1	4	7
1-1000	1	2	21	2	1	4	6

表8 "comp.ai*" データベース援用時の最頻出100語を除いた関連キーワード数の平均
Table 8 Average number of the derived keywords using the "comp.ai*" database except for commonly used 100 words.

頻出順位	K'_{c_1}	K'_{c_2}	K'_e	K'_d	K'_i	K'_h	K'_o
1-100	0	0	10	0	0	2	3
101-200	0	0	13	0	0	3	4
201-300	0	0	10	0	0	3	4
301-400	0	0	10	0	0	3	4
401-500	0	0	16	0	0	3	4
501-600	0	0	23	0	0	3	4
601-700	0	0	16	0	0	3	4
701-800	0	0	15	0	0	3	4
801-900	0	0	15	0	0	3	4
901-1000	0	0	13	0	0	3	4
1-1000	0	0	14	0	0	3	4

つデータベースから関連キーワードが導出されるため強い相関性を持ち，関連キーワードが持つサポート値が一般性が高い単語のサポート値よりも大きくなるため，平均的なキーワード空間からの相関ルール導出が抑制されることによる．

さらに，表5の $\mathbf{K_e}$ に対する表6の $\mathbf{K'_e}$ の値と表7の $\mathbf{K_e}$ に対する表8の $\mathbf{K'_e}$ の値を比べると，領域の異なるデータベースの援用により導出キーワード数の平均が下がるが，導出されるキーワード集合における最頻出100単語が占める割合 $(1-\mathbf{K'_e}/\mathbf{K_e})$ はともに30%程度と同程度である．これは，$\mathbf{K_e}$ がタイトルと著者の関連からアルゴリズム1を適用して拡大されたキーワード空間を導出に用いたため，著者を中心としたキーワード空間の偏りが発生しており，導出されるキーワード集合にも偏りが生じたためである．

アルゴリズム3により導出される関連キーワードの性質を調べるため，本稿で用いた3種類のデータベース (A, B, C) を援用した場合および援用しない場合

表9 異種データベースから得られる共通のキーワード数の平均
Table 9 Average number of the common derived keywords using the heterogeneous databases.

頻出順位	N∩A	N∩B	N∩C	A∩B	A∩C	B∩C
1-100	1	0	0	0	0	0
101-200	1	1	1	0	0	0
201-300	1	1	1	1	0	0
301-400	1	1	1	1	1	1
401-500	1	1	1	1	1	1
501-600	1	1	1	1	1	0
601-700	1	1	1	1	1	1
701-800	1	1	1	1	1	1
801-900	1	1	1	1	1	1
901-1000	1	1	1	1	1	1
1-1000	1	1	1	1	1	1

N：非使用，A：comp.ai*，B：sci.electronics*，C：sci.physics*

表10 異種データベースから得られる最頻出100語を除いた共通のキーワード数の平均
Table 10 Average number of the common derived keywords except for commonly used 100 words using the heterogeneous databases.

頻出順位	$(N\cap A)'$	$(N\cap B)'$	$(N\cap C)'$	$(A\cap B)'$	$(A\cap C)'$	$(B\cap C)'$
1-100	0	0	0	0	0	0
101-200	0	0	0	0	0	0
201-300	0	0	0	0	0	0
301-400	0	0	0	0	0	0
401-500	1	0	0	0	0	0
501-600	1	0	0	0	0	0
601-700	1	0	0	0	0	0
701-800	1	1	0	0	0	0
801-900	1	1	0	0	0	0
901-1000	1	1	0	0	0	0
1-1000	1	0	0	0	0	0

N：非使用，A：comp.ai*，B：sci.electronics*，C：sci.physics*

に導出されるキーワード集合 $\mathbf{K_o}$ に対して，すべての組合せの共通部分の単語数の平均を求めた結果が**表9**である．また，表9から最頻出の100単語を除いた結果が**表10**である．これらを見ると，導出されたキーワード集合には共通部分がほとんどない．つまり，アルゴリズム3により導出される関連キーワードは，用いるデータベースにより結果が異なっており，導出対象となるキーワード空間に偏りが発生している．一例として，11番目の出現回数4万回弱を持つ一般性が高い単語である "control" から導出されたキーワード集合を**表11**に示す．

以上より，アルゴリズム2により，検索対象とするデータベースの性質に応じたキーワードが導出されることが分かった．また，適切な領域のデータベースを選択することで，アルゴリズム3により特定の検索式の支援が可能であることが分かった．よって，本稿で提案したアルゴリズムが，文献データベース情報検索に対して有効であることが確認できた．

表 11 異種データベースから得られる関連キーワード
Table 11 Derived keywords using the heterogeneous database.

異種データベース	関連キーワード
N： 非使用	system, process, model, adaptive, time
A： comp.ai*	system, fuzzy, alberta, AI
B： sci.electronics*	system, process, model, ftp, armory, electronics, adaptive, power, rstevew, http, pub, www, circuit, remote, time
C： sci.physics*	system, fusion

7. 結　論

計算機環境の浸透にともない，電子図書館や電子出版などが注目されているが，基本的な文献情報に関わる実効性の高い検索システムの提供は依然困難である．また，現在，分類・主題分析などを行わず，キーワード統一を行わないフリーキーワード検索が増えており，現時点で抱える問題に対して効果の高いシステム設計指針を示すことは重要である．

本稿では，大規模データ処理システムの基礎的な解決技術となるデータマイニング技術を視野に入れ，文献情報検索において領域知識や背景知識が不足する場合に有効な検索支援を行う情報検索システムの構築を行った．重み付けを考慮した相関ルール導出アルゴリズムの拡張においても計算量は十分に抑制されており，リアルタイム性のある検索支援システムの実現が，現在稼働する計算機システムにおいて可能となった．また，文献データベースにおける属性を利用したキーワード空間の拡大により，効果的な関連キーワード導出を試みた．さらに，検索対象とする領域に偏りを持つデータベースを援用する検索支援の方法を示した．それらに基づいて，文献データベースに対する情報検索支援システムを実際に構築し，文献情報検索ユーザに効果的な検索式の改善方法を提供できることが示せた．今後，アルゴリズムの並列化，クラスタリングなどの異なるアルゴリズムの利用など，有効な支援システムの構成が可能なアルゴリズムを明確にしたうえで，現システムの提供する検索支援をより高度化していく必要がある．

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．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），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，本稿に対して貴重かつ有益なご指摘なら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) Etzioni, O.: The World-Wide Web: Quagmire or Gold Mine?, *Comm. ACM*, Vol.39, No.11, pp.65–68 (1996).
2) Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.: *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press (1996).
3) Feldman, R. and Dagan, I.: Knowledge Discovery in Textual Databases (KDT), *Proc. 1st Int'l Conf. on Knowledge Discovery and Data Mining (KDD-95)*, pp.112–117 (1995).
4) 伊藤耕一郎，河野浩之，長谷川利治：異種データベースからの相関ルールによる知識発見—WWW検索式の生成支援システムへの適用，第8回データ工学リークショップ (DEWS'97) (1997).
5) 川原　稔，河野浩之，長谷川利治：図書・文献データベースに対するナビゲータの構築，情報処理学会研究報告，97-DBS-112, pp.33–40 (1997).
6) 川原　稔，河野浩之：文献二次情報データベースにおける検索支援，情報処理学会研究報告，97-DBS-113, pp.239–244 (1997).
7) 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).
8) 河野浩之，長谷川利治：WWW情報空間における文書データマイニングを用いた知的検索システム，アドバンストデータベースシンポジウム ADBS'96, pp.27–34 (1996).
9) Lagus, K., Honkela, T., Kaski, S. and Kohonen, T.: Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration, *Proc. 2nd Int'l Conf. on Knowledge Discovery and Data Mining (KDD-96)*, pp.238–243 (1996).
10) Parsaye, K., Chignell, M., Khoshafian, S. and Wong, II.: *Intelligent Databases*, John Wiley & Sons, Inc. (1992).
11) Salton, G.: Another look at automatic text-retrieval system, *Comm. ACM*, Vol.29, pp.648–656 (1987).
12) Salton, G. and McGill, M.J.: *An Introduction to Modern Information Tutoring Systems: Lessons Learned*, McGraw-Hill, New York (1983).
13) 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).

14) Zaine, O.R. and Han, J.: Resource and Knowledge Discovery in Global Information Systems: A Preliminary Design and Experiment, *Proc. 1st Int'l Conf. on Knowledge Discovery and Data Mining* (*KDD-95*), pp.331–336 (1995).

(平成 9 年 9 月 1 日受付)
(平成 10 年 2 月 2 日採録)

川原　稔（正会員）

昭和 38 年生．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．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持つ．人工知能学会会員．

河野　浩之（正会員）

昭和 37 年生．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となる．同時に応用システム科学教室助手を兼任．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，現在に至る．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持つ．ACM，IEEE，AAAI，電子情報通信学会，人工知能学会各会員．

長谷川利治

昭和 9 年生．昭和 32 年 3 月大阪大学工学部通信工学科卒業．昭和 34 年 3 月同大学院修士課程修了．昭和 37 年 Johns Hopkins 大学より M.S. 修得．昭和 38 年 3 月大阪大学大学院工学研究科博士課程退学．同年 4 月同大学工学部通信工学教室助手となる．昭和 39 年ディジタル通信方式に関する研究により，大阪大学工学博士の学位を取得．昭和 40 年京都大学工学部数理工学教室助教授，昭和 47 年同教授となる．同大学工学部数理工学教室論理システム講座および同大学院応用システム科学専攻情報通信システム講座を担当．平成 8 年 4 月同大学院工学研究科応用システム科学専攻教授．情報伝送および処理，多値論理および回路，道路交通管制システムなどの研究に従事．日本 OR 学会，日本自動制御協会，IEEE 各会員．