カバー関係の抽出に基づく問い合わせ支援

劉 野[†1] 陳 漢雄[†2]
能登谷 淳一[†3] 大保信夫[†4]

文献データベース情報検索システムにおいて，ユーザは文献のキーワードの性質についての十分な知識を有していないため，最初から適当なキーワードを問合せとして与えることは困難である．ユーザの要求する全ての文献を洩れずに抽出するような問合せはしばしば多量の非関連文献を同時に抽出するため，システムによる修正支援が必要である．本研究では関連ルール (Association Rule) の手法を用い，キーワード間に存在する関係を発見してユーザの問合せ修正を支援するモデルを提案した．このモデルはStem Ruleとカバーの概念を統合し，サポートと信頼性の高い修正候補を排除する．これにより，従来の手法の問題点である，十分な絞り込みが困難である点，生成される修正候補数が膨大である点，修正問い合わせが元の問い合わせ結果を保証できなくなる点について解決を試みた．

Mining Coverages for Query Refinement

YE LIU,[†1] HANXIONG CHEN,[†2] JUNICHI NOTOYA[†3]
and NOBUO OHBO [†4]

In this paper we present a query support model for document retrieval system by mining Association Rule between keywords from large document databases. The model is proposed to normalize the structure of mined Association Rules. Two concepts, "stem rule" from which all other association rules can be delivered, and "coverage" which guarantees the retrieval result of the original query, are integrated into the model. These concepts, additionally with maximum support and maximum confidence, reduce the size of the rule base considerably, and enable the refined query to have a good recall. We build an interactive interface to aid user to refine their query. Empirical results confirm the screening effectiveness of our system.

1. まえがき

近年，電子図書館や電子出版などの普及によってオンライン文献が急激に増加している．そのような文献データベースの増大に伴い，それらを対象とした文献検索や知識発見がますます重要な課題になりつつある．

文献データベースに対する検索においては，適切ではない問い合わせは，大量の不関連文献を導出する．文献データベースにおいては，キーワードと関連文献の関係，キーワード間の関係は適切な問合せを作成する上で重要な知識であり，それらの知識を持たないユーザにとって，適切な問い合わせを発行することは困難と考えられる．そのため，文献データベースのユーザは，効率的検索のために検索対象となるデータベースに格納される文献データの性質に関する十分な知識を持つことを要求されている．しかしながら，大規模文献データベースに関して，必要とされる知識をユーザがデータベースから抽出し活用することは現実的ではない．そこで，問い合わせの作成に必要となる知識の文献データベースからの自動抽出と，それらの知識を用いたシステムによる問い合わせ作成支援が望まれる．

従来の一般的な問い合わせ作成支援に関する研究としては，問い合わせ拡張 (Query Expansion) やフィードバック (Relevance Feedback) などに関する研究などが存在する[2),6),15),17),18)]．

一方，大規模データベースからの知識の抽出に関しては，データマイニング (Data Mining) や知識発見 (Knowledge Discovery) の分野において，さまざまな手法が提案されている[1),3),4),9),10),14),16)]．

†1 アンリツエンジニアリング株式会社
Anritsu Engineering Co., Ltd.
†2 つくば国際大学産業情報学科
Institute of Industrial Information, Tsukuba International University
†3 筑波大学工学研究科
Doctoral Program in Engineering, University of Tsukuba
†4 筑波大学電子情報工学系
Institute of Electronics and Information Science, University of Tsukuba

文献データベースにおけるキーワード間の関係を対象とした知識発見[7),8),11)~13)]に関して，著者らは文献7)，8)，12)，13)において，キーワード間に存在する関係の関連ルール（Association Rule）としての抽出と，問い合わせの修正（Query Refinement）支援への利用を試みている．これらの手法では関連ルールの発見に際し，文献1)の手法に見られる最小サポート，最小信頼性の代わりに最大サポートと最大信頼性を用いることにより，検索結果を絞り込む形での問い合わせ修正支援を可能としている．

しかし，これらのしきい値を用いる関連ルール生成においては，システムによる支援の結果，修正を受けた後の問い合わせ（修正問い合わせ）が最初にユーザが発行した問い合わせ（初期問い合わせ）の検索結果を保証できないという問題が発生する．すなわち，初期問い合わせによる検索結果となる文献集合中で，どのような修正問い合わせの結果集合にも含まれないような文献が存在し，結果として，ユーザが真に要求している文献を検索することができない場合がある．また，これら文献データベースを対象としたルール生成手法の多くについて，関連ルールの構造などについての議論が十分に行われていないため，特定の応用分野に限定した場合でも有効性を示すことは困難であり，また，汎用性や拡張性に欠けるという問題がある．

本研究では，カバーの概念を導入し，従来の手法と同時に用いることにより，問い合わせの修正に関する各課題の解決を目指す．

- 効率よく絞り込むような修正候補の決定
- 提示する修正候補数の削減
- 初期問い合わせに対する検索結果の保証

次節では例を用いて，動機とアプローチの概要を説明する．3節では極小カバーを含む関連ルールによる問い合わせ修正のモデルを提案し，ルールの生成と極小カバーの生成アルゴリズムを記述する．4節では実験により有効性などについての評価を行う．

2. 本手法の概要

関連ルールによる問い合わせ修正のアプローチを図1に示す．検索に先だって，文献データベースからキーワードを抽出し，キーワードから関連ルールを生成し，ルールベースに格納する．検索時には関連ルールの適用による問い合わせ修正が行われる．ユーザの問い合わせqに対し，文献集合$D(q)$が検索されるが，問い合わせをq'に修正することにより，よりユーザの要求に近い，絞り込まれた文献集合$D(q')$を得る．本研究では，文献データベースの各文献からキーワードリストが抽出可能であるか，もしくは文献がキーワードリストを持つことを仮定する．以下では例を挙げて問い合わせの修正過程を説明する．

例1 文献データベース$\mathcal{D} = \{d_1, d_2, d_3, d_4, d_5\}$の各文献のキーワードリストが**表1**により与えられたとする．

$\mathcal{D}$に関する知識をあまり持たないユーザが潜在的に文献$\{d_3\}$を検索したいと考え，キーワードの集合$q = \{k_2\}$を問い合わせとして与えるとする．このとき，結果として$\mathcal{D}$の8割に相当する文献$\{d_1, d_2, d_3, d_5\}$が返される．一方，このケースにおいて理想的な問い合わせは明らかに$q' = \{k_2, k_3, k_4\}$である．従って，ユーザ問い合わせqに対する修正とはqからq'を導くことであり，これはqから連想されるキーワード集合$\{k_3, k_4\}$を見つけ，qに付け加える処理である．本研究のアプローチにおいて，「連想」はデータマイニングの手法で用いられる関連ルールが存在することと解釈できる．

しかし，qとキーワード集合$\{k_3, k_4\}$との間の関連ルール$q \Rightarrow \{k_3, k_4\}$の発見はしばしば非常に大きなコストを伴う．そこで，本研究ではqに対して，関連ルール$q \Rightarrow p_i$の適用により検索結果を効率よく絞り込むようなキーワード候補p_iのリストを提示し，ユーザがリストからキーワードkを選んで$q_1 \leftarrow q \cup \{k\}$のように問い合わせを修正するという，段階的な方法を採用する．このような問い合わせの修正の繰り返しによりユーザの要求に合致する問い合わせを得る．上の例で考えると，$q = \{k_2\} \rightarrow q_1 = \{k_2, k_3\} \rightarrow q' = \{k_2, k_3, k_4\}$が一つの修正過程の例であり，各修正段階で適用されるルール($q \Rightarrow \{k_3\}$や$q_1 \Rightarrow \{k_4\}$など)は後述のstem ruleである．

関連ルールはアイテム(Item)間のサポートと信頼性により生成される関係である[1)]．文献検索システムにおけるキーワードリストを持つ文献は，従来のデータマイニングにおけるアイテムリストを持つトランザクションに相当する．キーワード集合p_1とp_2間の関連ルールは3項組$(p_1 \Rightarrow p_2, \theta_s, \theta_c)$で表される．このルールは$p_1$を含んでいる文献が$\theta_s$のサポート(support, sptと略す)とθ_cの信頼性(confidence, cnfと略す)でp_2をも

表1 文献データベースの例
Table 1 An example of document base

$\mathcal{D}$	キーワードリスト				
d_1	k_1	k_2	k_3		
d_2	k_1	k_2			
d_3		k_2	k_3	k_4	
d_4				k_4	k_5
d_5	k_1	k_2			k_5

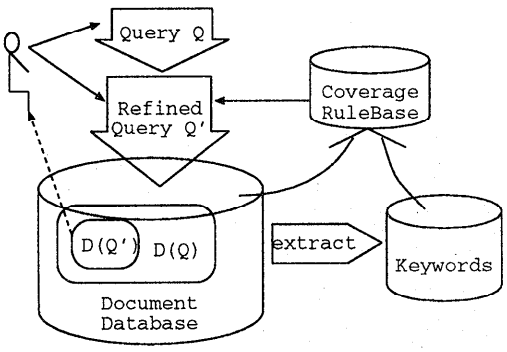


図1 問い合わせ修正の概要
Fig. 1 Overview

含むことを意味する．たとえば，5% の文献がキーワード k_1 と k_2 の両方を含んでおり，キーワード k_1 を含んでいる文献の中，30% の文献がキーワード k_2 を含んでいる場合は，$(\{k_1\} \Rightarrow \{k_2\}, 5, 30)$ と表す．

従来の問合せ修正の研究の多くにおいては，対称的な類似係数が利用されてきた．たとえばキーワードAとBがあるとき，AがBに80%「類似」していれば，BもAに80%「類似」していると考える．一方，関連ルールでは対称係数であるサポートと，非対称係数である信頼性が利用可能である．図2で4節の実験結果の一例を示す．この図において，根節は問い合わせに使われるキーワードとそのヒット数を示し，非根節は問合せの修正候補と修正された問合せのヒット数を示し，辺は関連ルールを示す．例えば，キーワード「画像通信」を含む文献を検索する，という問い合わせにヒットする文献は285件あり，この問い合わせにキーワード「マルチメディア」を追加するとヒットする文献数は108になる．このことはキーワード「画像通信」と「マルチメディア」がともに現れる文献の数は108，即ち，この2キーワードのサポートは108であることを示す．一方，図では示していないが，「マルチメディア」を含む文献数は737もあるため，(「画像通信」⇒「マルチメディア」)というルールの信頼性108/285は(「マルチメディア」⇒「画像通信」)というルールの信頼性108/737と異なり，結果の評価に差が生じる．このような，非対称係数の利用が，検索結果の絞り込みに有効であると考えられる．

2.1 検索結果を絞り込む関連ルールの生成

関連ルールを用いる従来のシステム[1),9),10),14),16)]では，最小サポートと最小信頼性を超えるものが適用すべきルールとして採用されてきた．文献検索においては，このように関連ルールを単純に適用した場合，以下のような問題が生じる．即ち，サポートや信頼性の高い関連ルールを適用して修正を行なっても検索結果のサイズが減少しない．

図2の左辺に示した，画像の転送に関する文献を検索する問い合わせを例として説明する．キーワード「画像通信」によるユーザ問い合わせを考える．ここで，「画像通信」を含む文献は285件存在する．これを絞り込む場合，関連キーワードの候補の例として「マルチメディア」，「ディジタル通信」などが考えられる．これらの候補から選ばれたキーワードによる絞り込みを考えると，候補「マルチメディア」のサポートは737と高く，これをユーザの元の「画像通信」に追加しても108のヒットが残り，絞り込みの効果は薄い．このとき，(「画像通信」⇒「マルチメディア」)というルールの信頼性は108/285とやはり高く，絞り込みの効果が薄いということがわかる．一方，候補の例「画像圧縮」と「ディジタル通信」を考えた場合，これらを最初の問合せに追加すると，出力の文献数は14に削減される．

以上のことを考え，本研究では最小サポートと最小信頼性を用いるよりもむしろ，最大サポートと最大信頼性を用いる方が検索結果の絞り込みに有効であると判断し，関連ルールの生成に最大サポートと最大信頼性を使用する．

2.2 構造化による提示候補数の減少

関連ルールを利用するためにはルールベースに管理が不可欠である．このとき，ルールベースのサイズが問題となる．また，ユーザに問い合わせ修正のためのキーワード候補を提示する際，大量の候補を提示するとユーザの混乱を招き，修正の意義が希薄になる．例えば，図2の左辺は問い合わせ「画像通信」を修正するためのキーワード候補の一部を表すが，1キーワードのみの候補でも800以上存在し，実用的ではない．

従来のアルゴリズムを用いてルールを生成した場合，キーワード集合 $\mathcal{K}$ に対して，理論的に $O(2^{|\mathcal{K}|})$ の組み合わせに対してチェックが必要である．経験的なしきい値を用いてルール数を若干削減することも可能であるが，以下に述べる候補のカバー問題を生じる．本研究では，ルールの構造やルール間の関係を解明する目的を兼ね，文献7)，13)で提案されるStem Ruleの方法により，ルールベースの縮小を図る．Stem Ruleの方法は，基本的な少数のルールのみを用いて，必要なルールを導出するための手法である．

2.3 カバー問題の解決

前述のように，問い合わせ修正のためのキーワード候補が多すぎる場合は，実用的ではない．このため，しきい値を超えるルールのみを関連ルールとして適用することにより修正候補数を減少する手法が2種類提案されている．一つは最小サポートと最小信頼性を用いる方法で

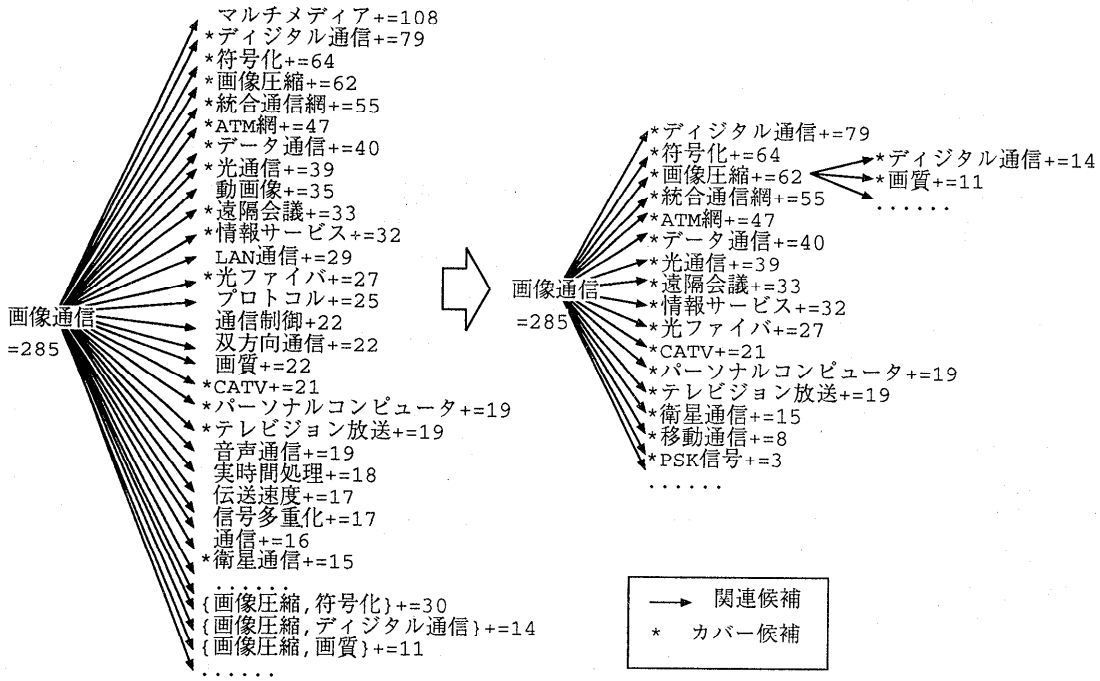


図2 問い合わせ修正の例
Fig. 2 Motivation

あり[1]，もう一つは最大サポートと最大信頼性を用いる方法である[7),12),13)]．しかし，これらの方法はキーワードの統計的なしきい値のみに着目し，キーワードと検索される文献とのセマンティクス上の関係を無視している．そのため，単純に一部カットされた修正候補を用いる問合せ修正では元の問い合わせの検索結果を保証できなくなる問題が生ずる．非形式的には，問い合わせ q により検索された文献がほかのキーワード集合 q' によっても検索できるとき，q' が q をカバーするという．しきい値を用いる場合，問合せ q に対して生成された修正 q' が q をカバーできない場合がある．極端な例では，q を含む文献のすべてのキーワードのサポートが最大サポートより大きい場合には，これらのキーワードは修正候補から排除される．したがって，q に対する修正は不可能となる．また，修正候補数の削減の要求に伴って，最小サポートのしきい値を上げるか，あるいは最大サポートのしきい値を下げるかのいずれか一方を選択しなくてはならないため，単純に排除される候補が増え，問題がさらに深刻になる．本研究ではこの問題を問い合わせの候補のカバー問題といい，この問題に対し，3節ではキーワード集合に対してはカバーを修正候補として生成し，構造化することにより解決を図る．

図2において，最小サポートを3, 5, 10にそれぞれ限定した場合，1キーワードのみからなる候補数は225, 124, 50にそれぞれ削減できる．しかし，最小サポートを3に限定した場合でも，既にもとの問い合わせ「画像通信」の検索結果は保証されず，最小サポートを10に限定した場合には，もと問い合わせの検索結果の7割の文献しか検索できない．一方，本研究では，問い合わせ候補間のカバー関係に基づいて，もとの問い合わせの検索結果を保証する極小候補集合を生成する．この極小候補集合は極小カバーと呼ばれる．カバーの関係に基づいて修正のプロセスは図2の右辺で示されたように構造化される．この図で一部示したように修正候補数は29に削減される．

3. カバー関係に基づく問い合わせ修正モデル

本節ではカバー関係に基づく問い合わせの修正モデルについて議論する．このモデルによって，問い合わせ，修正，関連ルールおよびカバーなどを統合的に扱い，修正の過程などの解明を試みる．

3.1 基本概念と記号の定義

文献検索システムにおける対象文献の集合とキーワードの集合をそれぞれ $\mathcal{D}$ と $\mathcal{K}$ として，文献 $d \in \mathcal{D}$ のキーワードを求める操作 $\rho : \mathcal{D} \to 2^{\mathcal{K}}$ を次のように定義する．

$$\rho(d) = \{k | k \in \mathcal{K} \text{ かつ } k \text{ が } d \text{ のキーワードである}\}$$

また，便宜上 $D \subset \mathcal{D}$ に対して，$\bigcap_{d \in D} \rho(d)$ のかわ

りに $\rho(D)$ と書く．

定義 1 （問い合わせと問い合わせ処理 σ） 問い合わせ q はキーワードの集合であり，即ち $q \subseteq \mathcal{K}$．問い合わせ処理 $\sigma : 2^{\mathcal{K}} \to 2^{\mathcal{D}}$ は次のように定義される．

$$\sigma(q) = \{\, d \mid d \in \mathcal{D}, \rho(d) \supseteq q\}.$$

即ち，問い合わせ q に対して，$\sigma(q)$ は $\mathcal{D}$ から q 中のキーワードを全て含むドキュメントを求める操作であり，$\sigma(q) = \cap_{k \in q} \sigma(\{k\})$ がつねに満たされる．

$q_1 \subset q_2 \subseteq \mathcal{K}$ のとき，$q_1 \Rightarrow q_2 - q_1$ は関連ルールという．このルールの信頼性 cnf とサポート spt については従来通り[1] の定義により，次の結果が得られる．

$$cnf(q_1 \Rightarrow q_2 - q_1) = \frac{|\sigma(q_2)|}{|\sigma(q_1)|}$$
$$spt(q_1 \Rightarrow q_2 - q_1) = |\sigma(q_2)|$$

また，カバーの形式的な定義は以下の通りである．

定義 2 （カバー） $\mathcal{D}$ は同上とする．$N \subset 2^{\mathcal{K}}$, $D \subset \mathcal{D}$ に対して，

$$D \subseteq \bigcup_{q \in N} \sigma(q)$$

が成り立つとき，N が D をカバーする (或は N が D のカバーである) という．また，N が D のカバーであり且つ $N' \subset N$ なる D のカバー N' が存在しないとき，N が D の極小カバーである．

同じように，次の式が成り立つとき，N が $N'(\subseteq \mathcal{N})$ をカバーするという．さらに，$N' = \{q\}$ のときは便宜上 N が q のカバーであるともいう．

$$\bigcup_{q \in N'} \sigma(q) \subseteq \bigcup_{q \in N} \sigma(q)$$

いうまでもなく，任意の与えられた D （もしくは N）の極小カバーは複数存在しうる．

3.2 SR ベース

定義 3 (SR ベース) SR ベースはグラフ

$$\mathcal{G} = (\mathcal{N}, \mathcal{E}, w, \gamma, C)$$

である．ただし，

$$\mathcal{N} = \{q \mid q \in \bigcup_{d \in \mathcal{D}} 2^{\rho(d)} \wedge (0 < w(q) \leq \theta_s)\}$$
$$\mathcal{E} = \{(q_1, q_2) \mid q_1, q_2 \in \mathcal{N} \ \wedge \ q_1 \subset q_2 \ \wedge 0 < w(e) \leq \theta_c)\}$$
$$w : \mathcal{N} \cup \mathcal{E} \to R^+$$

θ_s と θ_c はそれぞれ最大サポートと最大信頼性である．w はノードとエッジの重みであり，$q, q_1, q_2 \in \mathcal{N}, e = (q_1, q_2) \in \mathcal{E}$ に対して，w の定義は σ を用いて次のように表される．

$$w(q)) \stackrel{def}{=} |\sigma(q)|, \quad w(e) \stackrel{def}{=} \frac{|\sigma(q_2)|}{|\sigma(q_1)|} = \frac{w(q_2)}{w(q_1)}$$

γ はノードのペアからルールへのマッピングであり，$\gamma(q_1, q_2)$ は $q_1 \Rightarrow (q_2 - q_1)$ というルールを返す．w の定義から，ルール $\gamma(q_1, q_2)$ のサポートと信頼性はそれぞれノード q_2 とエッジ (q_1, q_2) の重みであることがわかる．

$C : \mathcal{N} \to \mathcal{N} \times 2^{\mathcal{N}}$ は下位カバーを表すが，詳しくは定義 4で説明される．

便宜上，$\gamma(q_1, q_2)$ がルールであるとき，q_1 が q_2 の上位ノード，あるいは q_2 が q_1 の下位ノードであるという．q_1 が q_2 の上位ノードであれば，$q_2 - q_1 \neq \emptyset$，定義 1により，

$$\begin{aligned}\sigma(q_2) &= \bigcap_{k \in q_2} \sigma(\{k\}) \\ &= (\bigcap_{k \in q_1} \sigma(\{k\})) \cap (\bigcap_{k \in q_2 - q_1} \sigma(\{k\})) \\ &\subseteq \bigcap_{k \in q_1} \sigma(\{k\}) \\ &= \sigma(q_1)\end{aligned}$$

ゆえに，任意のノードがその下位ノードをカバーする．

この性質により，ユーザは問合せノード q の下位ノードを選択し，検索結果のサイズが小さくなるように問合せを修正できる．ただし，この性質の逆は成り立たないため，q の複数の下位問合せの適当な和を選択してカバーを求めなければならない．

定義 2により定義されるカバーは本質的に二つの部分に分けられる．一つは正確に q しか含まない文献の集合 $\{d|d \in \mathcal{D} \wedge \rho(d) = q\} \stackrel{def}{=} \hat{D}_q$ のカバーであり，もう一つは残りの文献 $\sigma(q) - \hat{D}_q$ のカバーで，q の下位ノードにより構成されるカバーである．$\hat{D}_q$ のカバーに対してシステムはキーワードの追加によって文献集合の絞り込みができないため，$\mathcal{N}$ にあってもそれに関する修正ルールは生成されない．$\mathcal{N}$ のノードでもあるユーザ問い合わせの修正候補としては問い合わせノードの下位ノードによるカバーが用いられる．このようなカバーは以下のように定義される．

定義 4 （下位カバー） $K \subset \mathcal{K}$ が存在し，$C(q) = \cup_{p \in K}(q \cup p)$ が

$$\sigma(q) - \hat{D}_q$$

の極小カバーであるとき，$(\delta(q),\ C(q))$ を q の下位カバーとよぶ．ただし，

$$\delta(q) = \begin{cases} \emptyset, & \hat{D}_q = \emptyset \\ \{q\}, & \text{それ以外} \end{cases}$$

また，誤解が生じない限り $C(q)$ は差分 K だけに省略して $C(q) = K$ と書く．

例えば，表 1 の $\mathcal{D}$ に対して，$q = \{k_1, k_2\}$ とすると，$\sigma(q) = \{d_1, d_2, d_5\}$ であり，$C(q) = \{\{k_3\}, \{k_5\}\}$ は $\{d_1, d_5\}$ をカバーするが，$\hat{D}_q = \{d_2\}$ をカバーできるのは q 自身しかない．従って，q の下位カバーは $(\{q\}, \{\{k_3\}, \{k_5\}\})$ になる．

3.3 SR ベースの生成

我々のアプローチは基本的に $\mathcal{G}$ の定義に反映されている．$\mathcal{N}$ と $\mathcal{E}$ に対しては，それぞれ最大サポートと最大信頼性のしきい値を用いた制限を行っている．ルールは任意のノードのペアからではなく，エッジのみから生成され，このようなルールから全ての $\mathcal{G}$ のルールが導出可能である．文献 13) と同様に，このように生成されるルールを stem rule と呼ぶ．

本節は文献データベース $\mathcal{D}$ から $\mathcal{G}$ の条件を満足する SR ベースを生成するアルゴリズムについて述べる．比較のため，まず簡単に従来の手法[1),4)] を用いる場合の SR ベースの生成アルゴリズムを要約する．

(1) 最初に最小サポートと最小信頼性のしきい値を決める．

(2) キーワードの集合 $\mathcal{K}$ に対して，最小サポートを超えたキーワードからなる部分集合を 1-ItemSets とする．つぎに，1-ItemSets から任意の 2 つのキーワードのペアから，やはり最小サポートを超えたペアを集めて 2-ItemSets とする．同じように $|\mathcal{K}|$-ItemSets まで生成する．

(3) 各 Itemset に対して，その要素 q から $p \Rightarrow q-p$ をチェックし，信頼性が最小信頼性をこえたものをルールとする．

(4) 導出可能なルールを除去する．

(5) 各ノードに対して，カバーになる下位ノード以外のノードを除去する．

このようなアルゴリズムに従えば，n-Itemset の生成には ${}_{|\mathcal{K}|}C_n$ のキーワードの組合せをチェックしなければならず，全部の Itemset を生成するには $2^{|\mathcal{K}|}$ のキーワードの組合せのチェックが必要である．さらに，次のルール生成の段階では n-Itemset の 1 要素 q に対して同様に 2^n の ($q-p$ と p の) 組合せを調べるから n-Itemset に対して $2^n \times |n\text{-Itemset}|$ の組合せを調べなければならない．したがって，全ての Itemset に対しての計算量は以下の式になる．

$$\sum_{t=1}^{2^{|\mathcal{K}|}} (2^t \times |t\text{-Itemset}|) = O(2^{2^{|\mathcal{K}|}}) \qquad (1)$$

さらに，このアルゴリズムでは膨大な量の中間ルールを生成するだけでなく，問い合わせ候補のカバー問題を生ずる．これに対して，本研究で提案する以下のアルゴリズムでは問い合わせ候補のカバー問題に対処して，単一のキーワードから出発し，キーワードに対してサポートの高い順からカバーするまで stem rule のみを生成する．このヒューリスティックにより，少ないノードによってカバーが構成され，生成されるルール数が削減される．なお，本文中，特にアルゴリズムで使われる記号は**表 2** でリストした．

アルゴリズム (SR ベース $(\mathcal{N},\mathcal{E})$ の生成)
入力: $\mathcal{D}$，最大信頼性 θ_c
出力: $\mathcal{N}$, $\mathcal{E}$
$\mathcal{N}_0 = \emptyset, \mathcal{N} = \emptyset$, $\mathcal{E} = \emptyset$
/* 候補ノードの生成 */
forall $d \in \mathcal{D}$
begin
  $\rho(d)$ を求める
  forall $q \subset \rho(d)$
    $\mathcal{N}_0 = \mathcal{N}_0 \cup \{q\}$
end
/* $\mathcal{N},\mathcal{E}$ の生成 */
1 **while** $\mathcal{N}_0 \neq \emptyset$ **do**
begin
2 $|q| = \min_{p \in \mathcal{N}_0}\{|p|\}$ を満たす q を選ぶ
  $D = \sigma(q) - \hat{D}_q$
  /* q の下位カバー $(\delta(q), C(q))$ を生成 */
  $(\delta(q), C(q)) = \mathrm{MC}(q, D, \theta_c)$
3 $\mathcal{N}_0 = \mathcal{N}_0 - \{q\} - \{q \cup p | p \in \delta(q)\}$
4 $\mathcal{N} = \mathcal{N} \cup \{q\} \cup \{q \cup p | p \in \delta(q)\}$
5 $\mathcal{E} = \mathcal{E} \cup \{(q,\ q \cup p) \mid p \in C(q) \cup \delta(q)\}$
end

function $\mathrm{MC}(q, D, \theta_c)$
begin
  $C(q) = \emptyset$, $\delta(q) = \emptyset$
1 $q_0 = \{k \mid cnf(q \Rightarrow \{k\}) > 0\}$
2 **while** $D \neq \emptyset$ **do**
  begin
3 $|D \cap \sigma(\{k_0\})| = \mathrm{Max}_{k \in q_0}\{|D \cap \sigma(\{k\})|\}$
    を満たす k_0 を選ぶ．
4 **if** $cnf(q \Rightarrow \{k_0\}) \leq \theta_c$ **then**
      $C(q) = C(q) \cup \{\{k_0\}\}$
    else
    begin
5      $q' = q \cup \{k_0\}$

表 2 使用記号
Table 2 Notations

記号	意味	記号	意味
$\mathcal{D}$	全文献集合	$\mathcal{K}$	全キーワード集合
D	文献集合	K	キーワードの冪集合
q	問い合わせ	σ	問い合わせ処理
δ	$\hat{D}_q$ のカバー	ρ	キーワードを求める
$\hat{D}_q$	q しか含まない文献集合	$C(q)$	$\sigma(q) - \hat{D}_q$ のカバー

```
6          D' = D ∩ σ(q') − D̂_q'
7          if D̂_q' ≠ ∅ then
               δ(q) = δ(q) ∪ {{k0}}
           /* 再帰的に q' を展開 */
8          (δ(q'), C(q')) =MC(q', D', θc)
9          C(q) = C(q) ∪ C(q')
10         δ(q) = δ(q) ∪ δ(q')
       end
11     q0 = q0 − {k0}
12     D = D − σ({k0})
       /* 「極小」を保証するため，C(q) 中の
          k0 の下位ノードを削除する． */
13     C(q) = C(q) − {p | p ∈ C(q) ∧ p ⊃ k0}
   end
   return (δ(q), C(q))
end
```

定義により $\mathcal{G}$ のノードは最大信頼性を超えないことを条件としている．従って，アルゴリズムの中では，カバーを保証するため，最大信頼性を超えたノードを単純に切り捨てるのではなく「展開」する．

命題 1 上記のアルゴリズムは停止し，定義3の条件を満たす $\mathcal{G}$ を生成する．

証明. アルゴリズムにおいて，2行と3行を考慮すれば，MC の呼び出しで無限ループに陥らない限り，有限集合 $\mathcal{N}_0$ に対して1行の繰り返しは停止する．

一方，MC のなかでは1行と2行について

$q_0 \neq \emptyset \iff D \neq \emptyset$

を証明できる．従って，q_0 が有限であるため11行と合わせて2行の繰り返しが停止することも分かる．

また，アルゴリズムの5行と6行から

$(q_1, q_2) \in \mathcal{E} \iff q_2 \in C(q_1)$

であることが分かり，q の全ての下位ノードが q の下位カバーを構成する．また，リーフノード q の下位カバーは $(\{q\}, \emptyset)$ である．

また，アルゴリズムでは最初の $\mathcal{N}_0$ を生成する際，要素 q は $\mathcal{K}$ の代わりに $\rho(d)$ から選ぶ．言い替えると，$d \in \mathcal{D}$ を対象に計算するので，組合せのチェックは一文献の平均キーワード集合に対してのみ行なえばよい．$\mathcal{N}$ を生成するための計算量は

$$\sum_{d \in \mathcal{D}} 2^{|\rho(d)|} \approx |\mathcal{D}| \times 2^{|\rho(d)|}$$

になり，$\mathcal{N}$ の生成には $\mathcal{N}_0$ に対する組み合わせの必要がないため計算量が抑えられる．従って，$|\rho(d)| \ll |\mathcal{K}|$ を考慮すれば，式 (1) と比べると計算量が劇的に減ることがわかる．

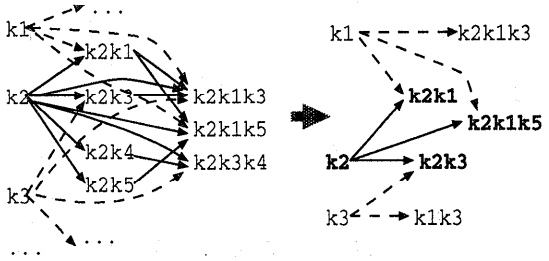


図3 カバー関連ルールの生成例
Fig. 3 Example of the Algorithm

以下では表1の文献データベースを用いてアルゴリズムを説明する．

$\theta_c = 0.5$ に対し $\{k_2\}$ のカバーを求めることは，$D = \sigma(\{k_2\}) = \{d_1, d_2, d_3, d_5\}$ のカバーを求めることと同等である．図3では $\mathcal{G}$ の生成を例示した．この図では $\{k_2\}$ のみを中心に実線で示している．左の部分は $\{k_2\}$ に関する全ての可能な上下位関係を表す部分グラフを示し，右の部分はアルゴリズムにより $\{k_2\}$ の下位カバーを示す (k2k1k3 は $\{k2, k1, k3\}$ の省略)．

まず，$q_0 = \{k_1, k_3, k_4, k_5\}$ から k_1 が選ばれ，$\{k_2\} \Rightarrow \{k_1\}$ の信頼性が $\frac{3}{4}$ であり，$\theta_c = 0.5$ を超えるため $\{k_2, k_1\}$ が展開される．即ち，再帰的に，

MC($\{k_2, k_1\}$, $D \cap \sigma(\{k_2, k_1\}) - \{d | \rho(d) = \{k_1, k_2\}\}$, 0.5)
=MC($\{k_2, k_1\}$, $\{d_1, d_5\}$, 0.5)

が呼び出され，結果として $C(\{k_2, k_1\}) = \{\{k_3\}, \{k_5\}\}$ が得られ，D が $D - \sigma(\{k_1\}) = \{d_3\}$ になる．

次に q_0 から k_3 が $\{d_3\}$ をカバーするので，$\{k_3\}$ が $C(q) = \{\{k_1, k_3\}, \{k_1, k_5\}\}$ に追加され，$\{k_3\}$ の下位ノード $\{k_1, k_3\}$ が $C(q)$ から削除される．このとき，D が空になるので，最終的に $C(\{k_2\}) = \{\{k_1, k_5\}, \{k_3\}\}$ となる．また，$\delta(\{k_2, k_1\}) = \{\{k_2, k_1\}\}$ は $\delta(\{k_2\})$ に合併され，$C(\{k_2\})$ とあわせて，アルゴリズムの4，5行より図3の右辺の k_2 に関する部分グラフが得られる．

4. 評　価

本研究で提案した手法の有効性を検証するため，問合せ支援のプロトタイプシステムを作成し，実験による評価を行った．

実験は電気工学分野の4万件の文献集合 ($|\mathcal{D}| = 40k$) を対象とした．文献集合全体に含まれるキーワード数は16717個である ($|\mathcal{K}| = 16717$)．

具体的な実験目的は次の2つである．

- ルールにより生成される問い合わせ修正のための候

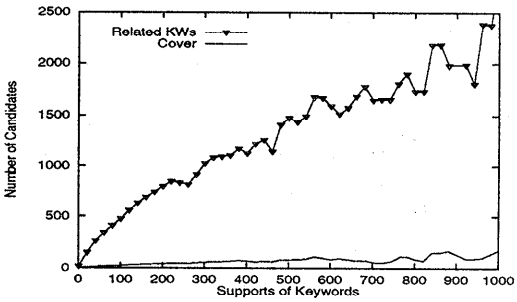


図4 キーワードの下位カバーと関連候補数
Fig. 4 Comparison with the number of relative keywords of a given keyword

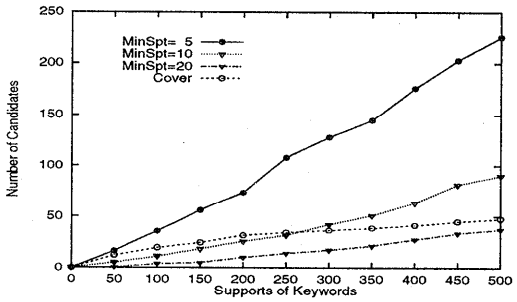


図5 最小サポート以上のキーワード候補数
Fig. 5 Number of relatvie keywords which larger than minimum supports

補キーワード数の削減効果の検証

- 修正問い合わせによる検索結果の正当性（初期問い合わせに対する保証）の検証

以下の実験結果の図中においては，本論文での提案手法による実験結果を，「Cover」というラベル名で示す．

4.1 修正候補の削減効果

まず，図4ではキーワードに対して本手法により生成される下位カバーのサイズ（つまり，問い合わせ修正に使われる候補数）を示した．図の横軸はキーワードのサポートを示す．図より，あるキーワードに対して下位カバーとなるキーワードは，そのキーワードの全ての関連キーワードと比較して，明らかに小さな量に抑えられていることがわかる．

同様に，図5では，最大信頼性0.6の時本手法により生成されるルールにより得られる候補キーワード数と，最小サポート (MinSpt) を5, 10, 20とし，最小サポートにより生成されるルールを制限した場合の候補キーワード数とを比較する．

図よりわかるように，最小サポートを5に設定した場合でも，全体的に候補数は本手法よりも非常に多い．最小サポートを10に設定した場合には，限定された区間において候補数が本手法による候補数をわずかに下回る．しかし，この区間以外では我々の方法より候補数が速く増加する傾向がみられる．さらに，最小サポートを用いた手法では，本質的にカバー問題が生じることについては後に図8に示す．

また，最大サポートを用いたルールの制限を行った場合の候補数は，図5と図4から以下の式により簡単に計算できる．

全関連キーワード数
= 最小サポートを s に設定時の関連キーワード数
+ 最大サポートを s に設定時の関連キーワード数

従って，図5と図4の縦軸に示される単位の差から，

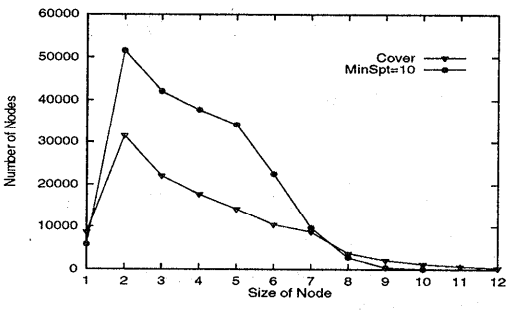


図6 最小サポート10のルールの候補数
Fig. 6 Number of rules with supports larger than 10.

最大サポートを用いる方法で生成される候補が膨大であることがわかる．

次に，図6では本手法と最小サポートを10と設定した場合の生成ルール数を比較する．横軸の目盛を従来の方法の n-Itemsets と一致させるためノードのサイズ n とし，縦軸は n-Itemsets の要素数に相当するようにした．関連ルールは Itemsets から生成されるため，カバーを用いることによって，より少ない関連ルールが生成される．特に，候補が集中する，小さいサイズのノードの区間においてその差が顕著に現れた．単純に MinSpt=10 によってルールを削減した場合でも，Itemsets に対して組合せを行なうため，ルールの数は46万前後であるのに対して，本手法では11万程度の stem rule しか生成しない．

また，図5では最大信頼性0.6の実験結果を示したが，図7では最大信頼性の値 (MaxCnf: θ_c) が本アルゴリズムに対する影響を調べた．この図，そして，図5との比較から分かるように，最大信頼性1.0, 0.6, 0.3と，開きが大きい値を設定しても生成されたキーワードの候補数はほとんど変わらない．

4.2 検索結果の評価

本研究では，本手法による支援下における問い合わせの修正のシミュレーションを行い，検索結果の評価を行

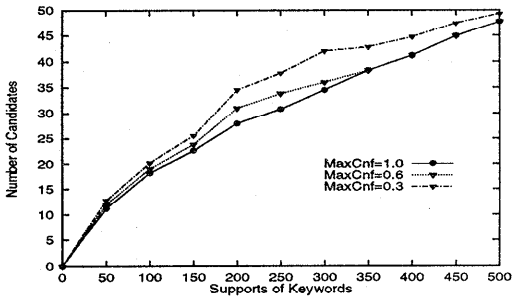


図7 最大信頼性の影響
Fig. 7 Effection of various MaxCnf's

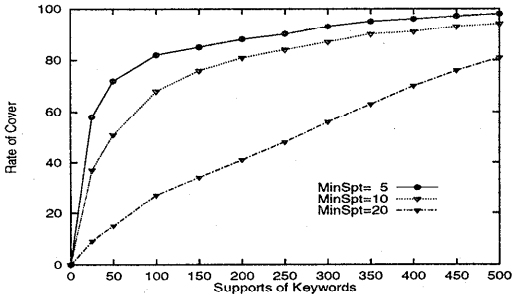


図8 候補のカバー率
Fig. 8 Rate of covering of refinement candidates

なった．シミュレーションの仮定として，ユーザはある文献の作者であるとし，文献に掲載される参考文献を作者の検索目標文献集合 (D_r) とする．本手法による問い合わせ修正の支援を利用して，この検索目標文献集合を検索し，検索結果を評価する．シミュレーションは次のステップで行なう．

- 検索目標集合の設定
  検索目標集合はある文献の最後に書かれた参考文献 D_r を仮定する．$D_t = D_r \cap \mathcal{D}$ は文献データベース $\mathcal{D}$ においての検索の目標集合である．
  例：*ACM SIGMOD'97* の予稿集からデータマイニングに関する論文の参考文献を目標集合とする．この場合，$D_r = 59$, $D_t = 30$ である．
- ユーザの問合せ Q $Max(D_t \cap \sigma(Q)$ により，D_t を最大にカバーするキーワードを問合せ Q とする．
  上の例において最初の問合せ Q は「データ発見」である．
- SR ベースを用いて，Q を修正する．
  例：問い合わせ Q=「データ発見」を次のような三つの問合せに修正可能である．
  Q_a=「データ発見」 AND 「知識獲得」 AND 「関連ルール」
  Q_b=「データ発見」 AND 「知識獲得」 OR 「データ発見」 AND 「関連ルー ル」
  Q_c=「データ発見」 AND 「知識獲得」 AND 「発見的方法」 OR 「データ発見」 AND 「関連ルール」 OR 「データ発見」 AND 「クラスタ」 AND アルゴリズム
- カバー率を計算する．

問い合わせのサポートとこの問合せに関するキーワード候補のカバー率の関係を図8に示す．この図では図4と同様に横軸はサポートを示す．図示のように，最小サポートによるルールの制限を行った場合，キーワード候補は問合せをカバーすることができない，即ち，ユーザが一部の文献を検索することが不可能になる．カバー率を上げるには，最小サポートのしきい値を下げる必要があるが，図5からも分かるように，最小サポートのしきい値の低下にしたがって生成されるルールの候補が著しく増大するため現実的ではない．これに対して，本手法では 100% カバーの条件下で，問合せを修正可能である．

5. まとめ

本論文ではキーワードによる文献検索に対するデータマイニングの応用として，カバーの概念を用いた問い合わせ修正モデルを提案し，関連ルールの生成と適用の方法について述べた．

本研究で提案するモデルはキーワード集合間のカバーの関係を用い，問い合わせ修正のプロセスに適した関連ルールの構造化を行う．モデルに従って作成された SR ベースによりルールを効率的に管理し，問い合せに対する柔軟な修正を支援するシステムを構築可能である．このモデルに基づいた問い合わせ修正効果を実験により検証した．

今後の検討課題として，アルゴリズムの効率面における改良と，シソーラスを参考にした，キーワード間の意味的関係の導入，などが考えられる．さらに，より客観的な評価を求めるため，文献検索の分野における標準である *TREC*[5] の提供するデータセットを用いた実験を行なう予定である．

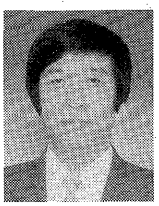
参考文献

1) Agarawal, R., Imielinski, T. and Swami, A.: Mining Association Rules between Sets of Items in Large Databases. *ACM SIGMOD'93*, pp.207-216 (1993)
2) Allan, J.: Relevance Feedback With Too Much Data. *ACM SIGIR'95*, pp.337-343 (1995)
3) Brin, S. *eg al*: Beyond Market Baskets: Generalizing Association Rules to Correlations. *ACM SIGMOD'97*, pp.265-276 (1997)

4) Brin, S. *et al*: Dynamic Itemset Counting and Implication Rules for Market Basket Data. *ACM SIGMOD'97*, pp.255-265 (1997)
5) Buckley, C. *et al*: Automatic query expansion using SMART : TREC 3. In D. K. Harman, editor, Overview of the 3rd Text REtrieval Conference. NIST Special Publication (1995)
6) Chen, C. M. and Roussopoulos, N.: Adaptive Selectivity Estimation Using Query Feedback. *ACM SIGMOD'94*, pp.161-172 (1994)
7) 陳漢雄，劉野，大保信夫: データマイニングのキーワード検索に対する応用. 情報処理学会研究報告, 97-DBS-113, pp.227-232. (1997)
8) Chen, H., Yu, J. X., Notoya, J., Liu, Y. and Ohbo, N.: A Data Mining Model for Query Refinement Revisite, *Proc. of IDEAL'98*, Springer-Verlag. pp.269-275. (1998)
9) Fayyad, U., Piatestsky, G. and Smyth, P.: From Data Mining to Knowledge Discovery in Databases. *3st Knowledge Discovery and Data Mining*, Califinia, USA, pp.37-53 (1996)
10) Han, J. and Fu, Y.: Discovery of Multiple-Level Association Rules from Large Databases *21st VLDB*, Zurich, Swizerland, pp.420-431 (1995)
11) 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(4), pp. 878-887 (1998)
12) Liu, Y., Chen, H., Yu, J. X.,and Ohbo, N.: A Data Mining Approach for Query Refinement. *PAKDD-98*, Australia. Also in *LNAI No. 1394*, Springer-Verlag, pp.394-396 (1998)
13) Liu, Y., Chen, H., Yu, J. X.,and Ohbo, N.: "Using Stem Rules to Refine Document Retrieval Queries". *Proc. 3rd FQAS*, Roskilde, Denmark. also in *LNAI No. 1495*, Springer-Verlag, pp. 249-260 (1998)
14) Savasere, A., Omiecinski, E. and Navathe, S.: An Efficient Algorithm for Mining Association Rules in Large Databases. *21st VLDB*, Zurich, Swizerland, pp.432-444 (1995)
15) Salton, G. and Buckley, C.: Improving Retrieval Performance By Relevance Feedback. *Jounal of The American Society for Information Science*, vol.41(4), pp.288-297 (1990)
16) Srikant, R. and Agrawal, R.: Mining Quantitative Association Rules in Large Relational Tables. *ACM SIGMOD'96*, pp.1-12 (1996)
17) Vélez, B., *et al*: Fast and Effective Query Refinement. *ACM SIGIR'97*, PA, USA, pp.6-15 (1997)
18) Xu, J. and Croft, W.: Query Expansion Using Local and Global Document Analysis. *ACM SIGIR'96*, pp.4-11 (1996)

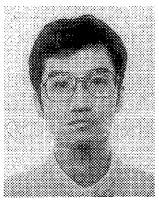
(平成10年9月20日受付)
(平成10年12月27日採録)

(担当編集委員 吉川 正俊)



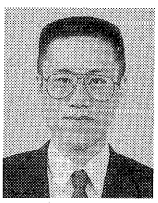
劉　　野（正会員）

1982年，中国東北大学自動制御科卒．1988年，東北電力大学大学院電子情報専攻修士課程了．同年9月同大学講師．1998年筑波大学工学研究科博士課程単位取得退学．同年アンリツエンジニアリング株式会社入社．情報検索，データマイニング，データベースシステムの研究に興味を持つ．



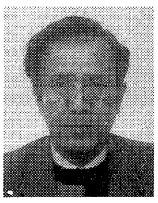
陳　　漢雄（正会員）

1993年筑波大学大学院博士課程工学研究科了．同年，同大電子・情報工学系助手．1994年，つくば国際大学産業情報学科講師．現在に至る．データベースシステムに関する研究に従事．工博．



能登谷淳一（学生会員）

1993年，筑波大学第三学群情報学類卒．1995年，同大大学院工学研究科博士前期課程了．現在，同大大学院工学研究科博士後期課程在学中．エンジニアリングデータベースシステムに興味を持つ．



大保　信夫（正会員）

1968年，東京大学理学部卒．1970年，同大大学院修士課程了．同年，同大理学部助手．1980年，筑波大学電子・情報工学系講師．1995年，同大電子・情報工学系教授．現在に至る．データベースシステムに関する研究に従事．理博．