データベースシステム 112－5
（1997. 5. 13）

図書·文献データベースに対するナビゲータの構築*

川原 稔† 河野 浩之‡ 長谷川 利治‡

†京都大学大型計算機センター ‡京都大学大学院工学研究科

図書·文献データベースに対する文献検索は，検索のために多くの領域知識や背景知識を要求されるため一般的に難しい．そこで，本稿では，図書·文献データベースの属性間の関連からキーワード空間の拡大を試み，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により導出されたキーワード集合を，文献検索に利用する手法について述べる．また，異種データベースからのルールを援用することによって，検索ユーザの要求に応じた検索を実現するための検索式の改善アルゴリズムも示す．なお，異種データベースに対する問い合わせを，エージェント通信の枠組を利用しながら行う文献検索支援ナビゲータのシステム構築を試みる．

キーワード：文献検索，データマイニング，相関ルール，異種データベース

Structure of the Navigation System using Association Rules in Bibliographic Databases

Minoru Kawahara† Hiroyuki Kawano‡ Toshiharu Hasegawa‡

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

Without background and domain knowledge, it is generally difficult for naive users to retrieve appropriate bibliographies from bibliographic databases. In this paper, in order to provide more helpful knowledge, we extend the mining association algorithms which discover relevance to a keywords space. By our algorithms, interesting rules are derived from relationships between several attributes in heterogeneous databases. We also propose algorithms in order to modify an initial query with keywords, which are specified by users' view. Moreover, we develop a navigation system using textual data mining algorithms and agent communication, and verify the effectiveness of our proposed algorithms.

Keywords: bibliographic search, data mining, association rule, heterogeneous database

* 連絡先: 〒606-01 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·文献データベースを用いた文献検索で，目的の文献を探し出すことが困難な場合が多いと指摘されており，多数の研究がなされている [11, 13]．実際，図書·文献データベースに対して数多くの検索式を投入しても目的とする文献が見つからず，熟練した図書館司書に頼んで文献を探し出して貰う場合も多い．これは，図書·文献データベースシステムを利用するに際して，検索対象分野に関する領域知識が必要であることに加えて，格納されているデータがどのような視点から作成されたのか，どのような特徴を備えたシステムを用いて構築されたのかといった背景知識も必要となるからである．

また，同様の問題は，SGML や，HTML による機械可読な文書データを大量に蓄積している WWW サーバ上のデータ検索においても生じている．むしろ，WWW は，文献データベースのように編纂組織が確立したものではないうえに，ネットワーク接続された膨大な数のホストによる緩やかな分散データベースとして構成されているため，問題をより困難にしている．そこで，フィルタリングにより特定の目的に沿った WWW ページを検索する研究 [1]，WWW や電子ニュースなどを異種データベース (heterogeneous database) と考えて概念階層を用いる検索手法の研究 [2] などが行われている．

そして，両システムに見られるように曖昧な位置付けをもつ文書データが著しく増加する状況で，検索に関わる問題を解決するためには，常に安定した処理が可能なアルゴリズムが基礎的な役割を担うと考えられる．この種の膨大なデータを扱う効率の良いアルゴリズムは，データマイニング (data mining) [4, 8, 15] に関わる分野において盛んに研究されており，実用性の高いルールを精度良く導くことが重要な目標になっている．また，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，様々な研究領域に関する枠組みが盛んに研究されている [5, 8]．例えば，文書データに対するアルゴリズムとしては，自己組織化マップ (SOM: Self-Organizing Map) によるクラスタ化 [9] や，テキストデータ発掘 (textual data mining) の研究 [6] が含まれる．

我々も，代表的なデータマイニングアルゴリズム [14] の拡張を試み，相関ルール (association rule) を利用した検索支援を行う RCAAU システムを開発し，テキストデータから導出されるルールの可能性を探っている [7, 8]．特に，検索ユーザが用いた初期キーワードを含むデータとの相関性が強いキーワード集合を求め，改善した検索式をユーザに提示する対話的検索の有効性を，“問答” と名付けた RCAAU システムにおいて検証している．

なお，本稿では，図書·文献データベースに焦点をあて，この種のデータベースに対する効果的な検索を実現するためのアルゴリズムを述べる．まず，複数のデータベースを効率良く検索するために，利用形態などによるデータベースアクセスに対する特徴を明確にする．また，タイトルや著者名を中心とした少数のデータから，より優れた相関ルールを抽出するための方法を述べる．さらに，WWW 情報空間，電子メール·電子ニュースなどから成る非対称な異種データベースを援用することによって，各種データベースから導出される相関ルールを基に，より検索ユーザの要求に応じた検索を実現するための検索式改善アルゴリズムについて述べる．加えて，エージェント通信の枠組を利用して，能力の異なるデータベースに対するデータマイニングを実行し，図書·文献データベースに対する検索ユーザの検索支援を行うナビゲータの構築を試みる．

以下，2 章では，図書·文献データベース検索の現状と検索の困難さについて簡単に考察する．3 章では，図書·文献データベースの検索ユーザが有効な検索を遂行する上で必要となるデータマイニングアルゴリズムを提案する．特に，異種データベースから求まるキーワード集合を用いた検索領域の拡大·縮小アルゴリズム，及び，相関ルールを用いた無意味語の削除アルゴリズムを中心に述べる．なお，4 章では，3章のアルゴリズムを用いた実装システムの状況等について述べ，5章に結論と今後の課題を述べる．

2 図書·文献検索システムの問題点

図書·文献データベースには，通常のデータベースと同様にデータベースの管理者がスキーマを設計し，データベース編纂者により正規化されたデータが格納されている．そのため，通常のデータベースは，入力される属性値の値域が制限されることが多く，検索ユーザが属性値を推測することは容易である．一方，図書·文献データベースでは，著者や出版社により属性の種類が異なることがある上に，文献データベース編纂者の分類方法によって属性·属性値も異なる．したがって，検索ユーザが用いるべき属性値を把握するのが，より困難となっていると言える．よって，通常のデータベースの利用者にくらべて，文献検索を行うユーザが目的のデータを得るのは難しいことが多い [11].

そこで，より優れた文献検索システム構築のために，文献情報に対する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 [11]．しかし，組織により索引付けの方法などが異なり，また，同一組織でも索引付けなどの方法を完全に統制することは，個人差があるため難しい．さらに，同一作業者による処理であってすら，時間的経過や体調などの要因により均一な索引付けは難しい [11].

また，各々の文書のもつベクトル空間に対する処理を行う検索 [12, 13] において，検索結果として求められる文章の質を，再現率·適合率を検索評価基準に用いて評価することもある．しかしながら，一般に文書ベクトル作成の手間が大きい上に，高い計算量をもつアルゴリズムが多く，大規模なデータに対して適用するには実用上困難な点が多いと考えられる．さらに，たとえ同一検索式が指定されたとしても，検索ユーザにより期待する検索結果も異なり，再現率·適合率を単純に評価基準とすることに対する議論が多いことにも注意を払うべきである．

なお，文献の電子化が進むにつれて，検索対象となる文書量が増加するだけではなく，背景知識や領域知識の不足した検索ユーザにより文献検索が行われる場合が，これまで以上に多くなっている．そのため，検索に関する知識の幅を広げる手法が必要であり，概念木 (conceptual tree)，分類 (taxonomy)，シソーラス (thesaurus) などの提供が考えられるが，この種の辞書作成にはコスト以外の問題も多い．なぜならば，一般に，組織や著者が異なれば単語に対する位置付けも異なり，関連性の強いとされるキーワードの質の良し悪しも異なるためである．したがって，概念木，分類，シソーラスを用いる単純な検索式の改善は，異なる観点のキーワードを混在させてしまうため，適切なデータの選択を難しくしてしまうと言える．このような点を含めて，キーワード空間の構造とその揺らぎの大きさの把握の難しさが，既存の典型的な文書データベースで用いられる AND，OR，NOT，NEAR などを用いた検索式を適切に与えることの困難さに繋がっている．

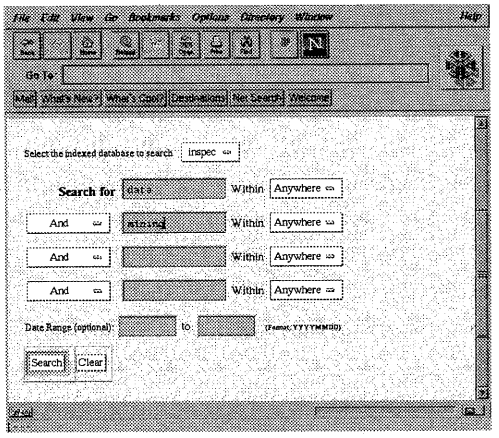


図 1: 現在の文献検索システムの例

しかしながら，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りつつあり，より大量の文献情報を蓄積するためには，ノイズ·検索漏れを防ぐ検索システムの必要性は非常に高い．そこで，本稿では，INSPEC データや雑誌記事索引データベースのデータを用いながら，図書·文献データベースに対して，より良い検索環境を提供するインターフェースを備えたナビゲータシステムの設計を試みる．なお，基礎的な検索式を処理する文献検索システムとして，全文検索システム (図 1) を併せて用いることにする．

3 異種データベースからのデータマイニングを用いた検索支援

本稿で構築を試みる図書·文献データベース検索システムは，単一のデータベースからの相関ルール導出だけでなく，ネットワーク上の各種データベースとの協調処理を目指す．また，キーワード空間に関する知識として，初期検索式に含まれるキーワードに関連するルールをデータベースから導出し提示する．なお，最小サポート閾値 ($minsup$) と，最小確信度閾値 ($minconf$) のヒューリスティックな設定も，良いルール導出のための重要な課題として扱う．

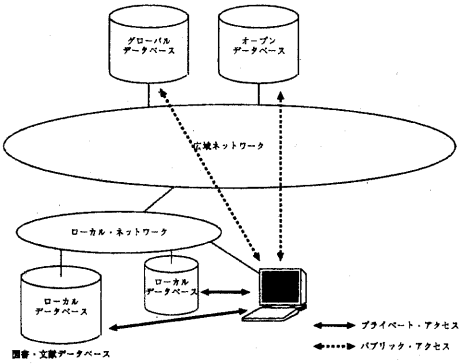


図 2: 非対称なデータベース情報空間

3.1 異種データベースの特徴

まず，キーワード空間を適切に拡大するために，異なる複数のデータベースの利用を考える．ただし，セキュリティや課金などに関連したアクセス制限，属性や属性値の異なるスキーマ設計，蓄積されているデータの質と規模など，大きな差異をもつ異種データベースとして存在することに注意を払わねばならない．そこで，この種の非対称なデータベース環境に存在する異種データベースの特徴を整理し，複数のデータベースに対する検索結果から得られる導出ルールを用い，検索ユーザの検索式に関わるキーワード空間を拡大して検索式の改善を行う．

まず，本稿の対象となる複数のデータベースは，図2に示した非対称なデータベース環境に存在する多様な情報システムを異種データベースとして扱うものであり，表1のように整理される．これらのデータベースの単語数とデータ量の特徴は，およそ表2のようにまとめることができる．

なお，独立して構築された各データベースは，複数のデータベースを相互に透過的に検索することを容易にするために，それぞれのシステムが共通のインターフェースをもつエージェント (agent) を実装し，異種データベース間の相違をエージェント層で吸収するものとする [7].

表 1: データベースの分類

分類	内容
ローカル	検索ユーザのプライベートなデータベースであり，ユーザの意思による部分的情報提供以外は，外部システムからのアクセスが制限される．
オープン	ネットワークを通じて公開されているパブリックアクセスが可能なデータベースである．ただし，検索結果のみを提示する．
グローバル	ネットワークを用いて公開されているパブリックアクセス可能なデータベースであり，データベース内に成立するルールを導出できる．単一もしくは複数のオープンデータベースの集合体に対してルール導出を行う機構を付加したシステムが，グローバルデータベースである．

表 2: データベースの規模

	単語数	データ量	例
図書·文献	小	大	国会図書館蔵書 INSPEC
ローカル	小	小	NetNews Email
グローバル	大	大	RCAAU
オープン	(小~) 大	(小~) 大	OpenText AltaVista

3.2 図書·文献データベースにおけるキーワード空間の拡大

図書·文献データベースには，中にはアブストラクトを属性としてもち，当該文献に関連する単語を多くもつ場合もあるが，一般的には，タイトルや著者などのデータをもつにすぎず，属性値に含まれる単語数は少ない．したがって，各属性に対して相関ルール導出アルゴリズムを用いて相関ルールを求めたとしても，相関性の高いキーワードは多くならない．

そこで，あらかじめシステムに対して関連する属性を指定し，それぞれに含まれる単語の直積集合を用いてクラスタを形成し，そのクラスタから得られる導出ルールによりキーワード空間を拡大することが考えられる (図3)．これは，例えば，図3において，属性 A_1 を“タイトル”，属性 A_2 を“著者”と見なすと，同一著者の著書のタイトル集合からキーワード空間を構成することに相当する．

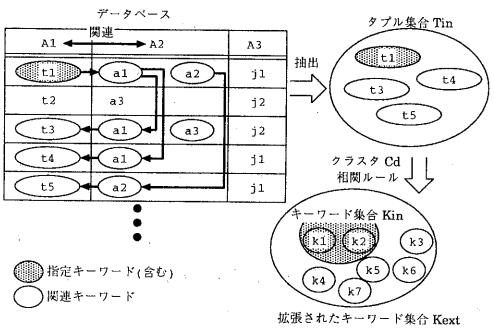


図 3: キーワード空間の拡大

キーワード空間拡大アルゴリズムは，次の通りである．

アルゴリズム

入力: 初期検索式におけるキーワード集合 $\mathbf{K_{in}}$

出力: 拡大されたキーワード集合 $\mathbf{K_{ext}}$

手順:
1. 指定された属性が，指定したキーワード集合 $\mathbf{K_{in}}$ に合致するタプル集合 $\mathbf{T_{in}}$ を抽出する．
2. $\mathbf{T_{in}}$ から，指定された属性に関連付けされた属性によりキーワードを抽出してクラスタ $\mathbf{C_d}$ を生成する．
3. クラスタ $\mathbf{C_d}$ に相関ルール導出アルゴリズムを適用して拡大されたキーワード集合 $\mathbf{K_{ext}}$ を導出する． □

3.3 異種データベースを用いた検索式生成アルゴリズム

3.2節で述べたアルゴリズムを，図書·文献データベースに適用して相関ルールを導出し，キーワード集合 $\mathbf{K_d}$ を導出する．検索対象である図書·文献データベースから得られるキーワード集合 $\mathbf{K_d}$ は，検索によるデータの存在が保証されており，また，属性間の関係を考慮するものであるため重要度が高い．しかしながら，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をもつため，そのキーワード空間は狭い．

そこで，グローバルデータベースおよびローカルデータベースから相関ルールを導出し，それぞれキーワード集合 $\mathbf{K_g}$ ならびに $\mathbf{K_l}$ を導出し，キーワード空間を拡大する．異種データベースから得られるキーワード集合は，検索式におけるキーワード空間を拡大して検索式の改善の可能性を与えるが，その適用方法は各々について注意深い考察が必要である．

まず，グローバルデータベースから得られるキーワード集合 $\mathbf{K_g}$ は，パブリックアクセスを許すデータから構成されているデータベースである性質上，一般的に成立するルールから求められたものが多い．ただし，一般的なルールは，検索ユーザの関心があるキーワード空間と隣接しているとは限らないことが多い．一方，ローカルデータベースから得られるキーワード集合 $\mathbf{K_l}$ 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，他のデータベースには適用できないキーワードも多い．

以上のことを考慮して，各キーワード集合を用いたアルゴリズムを提案する．

アルゴリズム

入力: 初期検索式におけるキーワード集合 $\mathbf{K_{in}}$
図書·文献データベース D_d
ローカルデータベース D_l
グローバルデータベース D_g

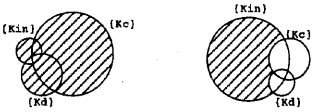
出力: 改善された検索式 E_o

手順:

1. 図書·文献データベース D_d から導出されるキーワード集合 $\mathbf{K_{ext}}$ において，導出ルールからキーワード集合 $\mathbf{K_d}$ を得る．
2. ローカルデータベース D_l における導出ルールからキーワード集合 $\mathbf{K_l}$ を得る．
3. グローバルデータベース D_g の導出ルールからキーワード集合 $\mathbf{K_g}$ を得る．
4. 上記ステップ 2 および 3 を，$\mathbf{K_l}$ と $\mathbf{K_g}$ の空でない共通部分集合 $\mathbf{K_c}$ 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 $minconf$ を緩和しながら繰り返す．
5. 論理和 $\mathbf{K_{in}} \cup \mathbf{K_d} \cup \mathbf{K_c}$ を取り，ローカルデータベースから導出された残りのキーワード $\mathbf{K_l} - \mathbf{K_d} - \mathbf{K_c}$ を表示順の重み付けとした改善された検索式 E_o を出力する． □

なお，本稿の冒頭に述べた $minsup$ と $minconf$ の閾値を適切に設定するために，ステップ 4では，$minsup$ および $minconf$ をエージェント間で交換しながら緩和する．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，エージェント間通信による緩和手法によって，有効な情報をなるべく失うこと無く，多くの関連キーワード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．

また，ステップ 5 では，キーワード集合の改善方法として，単純に OR を取るものとした．しかし，キーワードの特徴に応じて，以下に述べるような改善方法を動的に選択できるように考慮しておくべきである．例えば，図 4 のように，検索ユーザが入力したキーワード集合 $\mathbf{K_{in}}$ が小さい場合には OR 演算によるキーワード集合の拡大が有効である可能性が高く，大きい場合には NOT 演算によるキーワード集合の縮小が有効である可能性がある．さらに，この様な OR 演算や NOT 演算を用いた検索式に適合した検索領域から相関ルールを求めることも有効であると考えられる．なお，相関ルールにより関連キーワードを求めた場合，無意味語 (stop words) が含まれることがあり，OR 演算で領域を拡大する際には，無意味語の除去も効果的に行う必要がある．



(a) OR を好む場合 (b) NOT を好む場合

図 4: キーワードによる被覆領域の拡大と縮小

3.4 相関ルールによる無意味語の除去

検索キーワードに対する関連キーワードを相関ルールから求める場合，多くのキーワードに関して成立する相関ルールは，無意味語に関わるルールとなりがちであり，有効なルールを効果的に選択する必要がある．これは，文書全体に含まれる単語を重み付けせずに均一にデータベース化しなければならない状況で生じることが多く，“of”, “the”, “and” などが該当する．

そこで，我々は，n 個の領域の異なる複数のデータベースのキーワード集合 $\mathbf{K_i}(1 \leq i \leq n)$ に高い頻度で存在しているキーワード k_j を，異なるデータベースから導出されるルールを削除する最小の領域数 $mincommon$ を閾値として与えることによって取り除くこととした [7]．これにより，単一データベースのキーワード頻度のみによらずに，無意味語を効果的に削除した相関ルール導出が可能となった．

4 相関ルールを用いたナビゲータの構成

本章では，図書·文献データベースとして，国立国会図書館の雑誌記事索引データベース，および，INSPEC データベースを用いて，ナビゲータシステムを構成する．

まず，国立国会図書館の雑誌記事索引データベースは，約 6,000 誌の雑誌の論文に関して年間約 230,000 件，人文·社会·科学関係の全分野のデータが入力されたものであり，最近の 79,244 件 (キーワード数 96,212 語) のデータ

を用いた．また，INSPEC データベースは，英国 INSPEC が文献の収集･整理を行な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，最近の 3,353,605 件 (キーワード数 429,316 語) のデータを格納している．なお，両データに対する全文検索は，全文検索システム OpenText によって実現している．

なお，これらのデータベースの内，雑誌記事索引データベースに対して，タイトルのキーワードを解析したデータベースを構成し相関ルールを導出した．図 5と図 6は，検索語として “環境” を与えた実行例である．検索された文献数は，全文検索システムが 268 件なのに対して本システムでは 1,941 件に及んでおり，関連語としても，

> 特集，問題，地球，開発，エネルギー，
> 都市，企業，教育，技術

が提示されており，絞り込み，または，関連語への移行の可能性を与えている．

さらに，グローバルデータベースである RCAAU を用いた場合，380,683 件の URL を解析して総キーワード数 366,041 語をもつ．図 7は，検索語として “環境” を与えた実行例であり，

> environment, 計算機，設定，動作，問
> 題，開発，学習，公害，教育

など関連性の高いキーワードとして上記と異なる語も含めて提示される．また，電子ニュースグループの記事合計約 120 MB の 16,000 通の記事と， 3 種のメーリングリストの合計約 7 MB の 1,000 通の電子メールから成るローカルデータベースも用いる [7]．なお，総キーワード数は 126,459 語であり，グローバルデータベースとの共通キーワード数は 1,073 語である．

以上，これらのシステムを連携させ，グローバルデータベースとローカルデータベースから導出される相関ルールを併せて用いることにより，検索式が改善され，より好ましい全文検索も可能となる．さらに，このようなナビゲータは，検索ユーザに対して改善した問い合わせ記述を提示するだけではなく，キーワード空間の構造と，入力キーワードが一般にもつ揺らぎを把握する機会を高めるものともなっている．

図 5: 全文検索システムでの検索結果画面

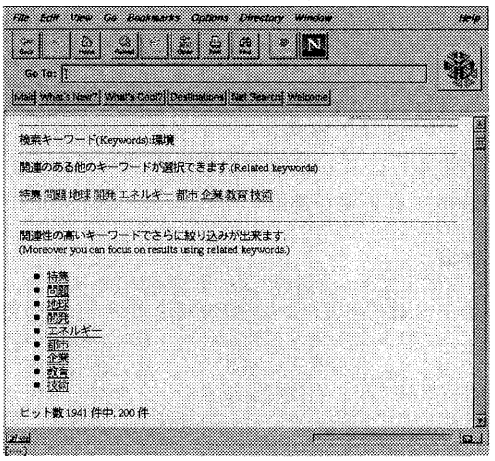


図 6: 本システムでの検索結果画面

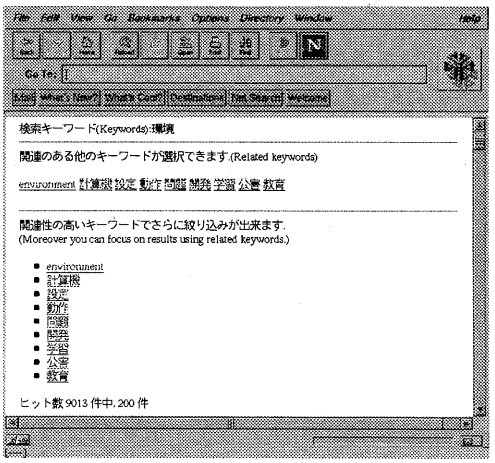


図 7: “問答” での検索結果画面

5 結論と今後の課題

コンピュータ環境の変化により，電子図書館や電子出版などが注目されているが，文献情報に対する有効な検索手段の提供は依然困難な状況である．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も重要である．

そこで，本稿では，キーワードが少なく検索が困難な文献情報検索において，領域知識が不足する場合に有効な検索の実現を，データマイニング技術を基礎に試みるナビゲータの構築を行った．なお，アルゴリズムの計算量を十分に抑制することによって，リアルタイム性の高い検索システムの実現が，現在稼働する計算機システムにおいて可能となると考えられる．

今後，より検索ユーザの自然な検索をサポートするために，検索式生成アルゴリズムをより洗練する必要がある．

謝辞

本システム構築のため，全文検索システム OpenText の試用提供および技術支援を頂いた日商岩井インフォコムシステムズ株式会社，また，新須哲朗氏，土屋悟氏，花房寛氏に感謝する．国立国会図書館より，本稿で用いた雑誌記事索引データベースのデータ部分を提供頂いたことに感謝する．最後に，本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] M. Balabanovic, Y. Shoham and Y. Yun, "An Adaptive Agent for Automated Web Browsing," Stanford University Digital Library Project Working Paper SIDL-WP-1995-0023, Stanford, 1995.

[2] M. Q. W. Baldonado and T. Winograd, "Techniques and Tools for Making Sense out of Heterogeneous Search Service Results," Stanford University Digital Library Project Working Paper, Stanford, 1996.

[3] S. Dao and B. Perry, "Applying a Data Miner to Heterogeneous Schema Integration," Proc. 1st Int'l Conf. on Knowledge Discovery and Data Mining (KDD-95), pp. 63–68, 1995.

[4] O. Etzioni, "The World-Wide Web: Quagmire or Gold Mine?," Communications of the ACM, Vol. 39, No. 11, pp. 65–68, 1996.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining," AAAI/MIT Press, 1996.

[6] R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases(KDT)," Proc. 1st Int'l Conf. on Knowledge Discovery and Data Mining (KDD-95), pp.112–117, 1995.

[7] 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 回データ工学ワークショップ (DEWS'97), 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] K. Lagus, T. Honkela, S. Kaski and T. Kohonen, "Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration," Proc. 2nd Int'l Conf. on Knowledge Discovery & Data Mining (KDD-96), pp. 238–243, 1996.

[10] J. Mayfield, Y. Labrou and T. Finin, "Evaluation of KQML as an Agent Communication Language," Proc. of the 1995 Workshop on Agent Theories, Architectures and Languages, Springer–Verlag, 1996.

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

[12] G. Salton and M. J. McGill, "An Introduction to Modern Information Tutoring Systems: Lessons Learned," New York: McGraw-Hill, 1983.

[13] G. Salton, "Another look at automatic text-retrieval system," Communications of the ACM, Vol. 29, pp. 648–656, 1987.

[14] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," Proc. of the 21st VLDB, U. Dayal, P. M. D. Gray and S. Nishio (Eds.), Zurich, Switzerland, pp. 407–419, 1995.

[15] O. R. Zaine and J. Han, "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.