情 報 学 基 礎 46－7
（1997. 7. 24）

雑誌記事索引データベースのナビゲーション －「問答」による分析－*

河野 浩之† 川原 稔‡

†京都大学大学院工学研究科 ‡京都大学大型計算機センター

雑誌記事索引データベースなどをはじめとする図書・文献データベースを用いた文献検索は，より困難になりつつある．これは，適切な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である．本稿では，データマイニング分野で盛んに研究されている相関ルールによりキーワード集合を求めるアルゴリズムを用い，雑誌記事索引データベースの属性値間の相関ルールを用いて連想的な検索キーワードとしてユーザにフィードバックするシステム構成について考察する．また，この種のアルゴリズムが，文献検索においてどの程度の効果をもつかを調べるために，RCAAU システムの一部の仕様を変更することによって実装を行い，その実験過程を示しながら議論を行う．

キーワード：文献検索，データマイニング，相関ルール，連想検索

Navigating System for Bibliographic Databases: Search Interface by Mondou

Hiroyuki Kawano† Minoru Kawahara‡

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

Without background and domain knowledge, it is generally difficult for naive search users to retrieve appropriate bibliographies from bibliographic databases. In this paper, in order to provide more helpful knowledge including associative keywords, we extend the algorithm to derive association rules, which plays very important roles in the area of data mining researches. Moreover, we develop a navigation system using textual data mining algorithms, and verify the effectiveness of our proposed algorithms. We also consider the effectiveness of data mining for bibliographic search.

Keywords: bibliographic search, data mining, association rule, associative search

* 連絡先: 〒606-01 京都大学大学院工学研究科応用システム科学専攻 河野 浩之
Tel (075)753-5513, Fax (075)761-2437, e-mail: kawano@kuamp.kyoto-u.ac.jp

1 はじめに

雑誌記事索引データベースなどをはじめとする図書・文献データベースを用いた文献検索で，目的の文献の検索が困難な場合が多いことが指摘されており，多数の研究がなされている [12, 14]．これは，データベース利用に際して，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，格納されたデータがどのように作成されたかの視点を含めて，システム構築に関連する特性などの背景知識が必要となるためである．

なお，同種の検索に関する問題は，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 Web サーバ上のデータ検索においても生じている [4, 9, 16]．例えば，フィルタリングにより特定の目的に沿った Web ページを検索する研究 [2]，Web 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 [3, 7] などが行われている．ただし，Web では，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おり，さらに，文献データベースのように編纂組織が確立していないために，問題がより困難になりがちである．

以上，これらの種類の問題点は，生データと考えられるような文書データが著しく増加する状況によって生じているものであり，検索処理に関わる問題を効果的に解決するには，常に安定した処理が可能なスケーラビリティの高いアルゴリズムが重要な役割を担うと考えられる．よって，生データを含む膨大なデータを効率良く扱うアルゴリズムに関わる，データマイニング (data mining) やデータベースからの知識発見 (KDD: Knowledge Discovery in Database) が非常に大きな役割を果たすものと期待される [5, 9]．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 [11] や，テキストデータ発掘 (textual data mining) の研究 [6] も含まれる．

そこで，本稿では，雑誌記事索引データベースによる図書・文献検索に焦点をあて，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に代表的なデータマイニングアルゴリズム [15] の拡張を行う．さらに，テキストデータから導出されるルールの可能性を探るために，雑誌記事索引検索システムの実装を試みるが，その際に，拡張した相関ルール (association rule) を利用して検索支援を行う "問答" と名付けた RCAAU システムを基盤とする [8, 7, 9]．なお，実装システムを用いて，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について考察する．

以下，2 章では，図書・文献検索システムの現状から検索の困難さについて簡単な考察を行う．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について述べる．さらに，4 章で，3章のアルゴリズムを用いた実装システムの現在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

一般に，データベース管理者がスキーマ設計を行ったデータベースに対して，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る．従って，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．しかし，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・属性値が異なる場合も多く存在する．よって，通常のデータベース利用に比べて，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．

そこで，検索に関するユーザの知識を広げる手法として，ユーザが容易に利用し得る概念木 (conceptual tree)，分類 (taxonomy)，シソーラス (thesaurus) などの提供が考えられているが，この種の辞書作成にはコスト以外の問題点がある．それは，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が異なるからである．つまり，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場合がある訳である．

そこで，より優れた文献検索システムを構築するために，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研究もなされている．しかし，やはり，組織による索引付けの方法の差，同一組織内でも索引付けなどを統制することの困難さ，加えて，個人差もあるため問題解決に対する指針にとどまりがちである [12]．

さらに，各文書のもつ文書ベクトル処理を行う場合 [13, 14]，検索結果の文章集合の質を，再現率・適合率を検索評価基準に用いて評価することがある．しかし，一般に文書ベクトル作成コストは大きく，非常に高い計算量をもつアルゴリズムが採用されがちであり，大規模データに対して適用するには実用上困難な点も多い．また，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注意すべきでもある．

しかしながら，文献情報の電子化は急速に進んでおり，検索対象となる文書量が増加するだけでなく，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機会が，これまで以上に増えている．そこで，上述したような様々な要因によってキーワード空間の構造とその揺らぎの大きさの把握は難しくなっているものの，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を何らかの形で解決する糸口が必要がある．つまり，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るが，より大量の文献情報を蓄積するために，ノイズ・検索漏れを防ぐ対話性の高い知的な検索システムが必要であると言える．

そこで，本稿では，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，より優れた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ション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は，図1に示した全文検索システムが併用できるものとする．

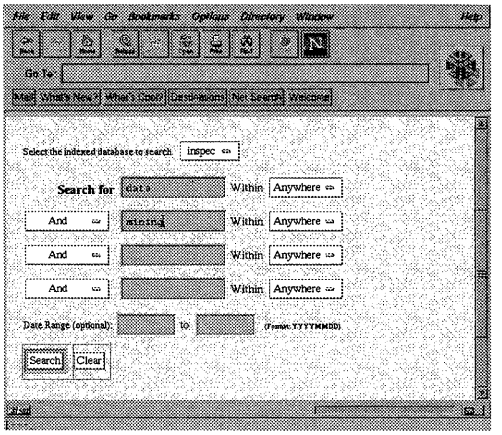

図 1: 全文検索システムの一例

3 検索支援アルゴリズム

本章では，広域ネットワークにおける協調作業の基盤となる検索支援を行う「問答」(RCAAU) の開発基盤 [8, 9, 7] をもとに，関連語・連想語検索システムに関わる実験を行う上で必要なアルゴリズムについて述べる．なお，基礎とする RCAAU の URL は，"http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/" である．

3.1 文書データマイニング

我々は，多様なデータ構造をもつデータベースに対してデータマイニングを行い，検索対象であるデータ集合の特性を把握する研究を進めている [8]．なお，本節では，特に文書データマイニング (textual data mining) に対して焦点を当て，知的検索支援を実現する技術について述べる．

まず，文書データに関する知識を導出するために，対象となる文書集合に対する考察を行う．

検索対象となる文書データは，複数の組織において独立に作成されていることから，同義語などを数多く使用した非均一な質をもつ文書であると考えられる．また，当然ではあるが，タイトル作成は自由に可能であるという特性から，情報空間全体で使用されるキーワード空間への制約を課すことは不可能である．このことにより，正確な検索記述を書くことは全く困難なことを示すものであると言える．

加えて，従来の多くのデータベースに格納されるデータは，データベース設計者などのもつ背景知識を用いたデータモデルが与えられているため，検索者がある程度の背景知識を共有すれば，適切なキーワードを用いた精度の良い検索式の記述がある程度可能であった．しかし，実際，データの存在の有無などの検索対象に関する知識を，ある程度ユーザがもつと考えられる，新聞記事などを対象とした文書検索ですら，精度の高い検索式は複雑な記述を必要とする．

従って，本稿で述べるシステムは，入力キーワードを用いた検索によって多量の検索結果を与えるだけでなく，有用度の高い情報として品質の高いキーワードを効果的に提示することが重要と考えられる．

そこで，入力キーワードに関係するルールをデータマイニング処理によって求め，ユーザの検索要求の近傍のキーワード空間の構造を関連語の知識として与える検索過程を有用と考える．

例えば，図 2に示したデータベース $\mathcal{D}$において，あるキーワード集合を用いた検索記述式によってデータ集合 $\{D\}$ が検索結果として求められる．この場合は，検索結果として非常に大きなデータ集合 $\{D\}$ をユーザに対して直接提示するのではなく，その部分集合である $\{d\}$ とともにルール集合 $\{r\}$ を検索結果として与えることとなる．

なお，キーワード空間の構造だけでなく，参考・引用文献を含めたハイパーテキストとして書誌データが構築されていれば，リンク構造の到達性を利用した n ステップ到達性を保証するデータ集合を示すなどの処理も考えられる [10].

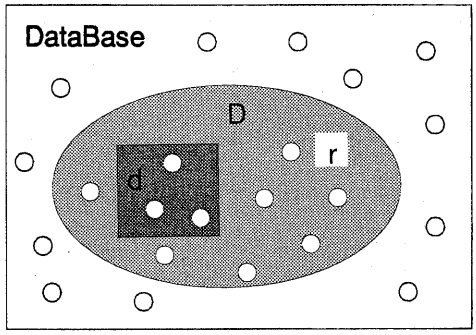

図 2: 検索結果と表示されるデータの関係

3.2 重み付き相関ルール導出アルゴリズム

本節では，多くの研究がなされている組指向アルゴリズム (tuple-oriented algorithm) の一つである相関ルール導出アルゴリズム [1] の概略について示し，検索に用いるキーワード集合に対して重みを与える拡張について述べる．

まず，相関ルール導出アルゴリズムの検索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 K_gと，導出されるキーワード集合 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 K_dへと加える．さらに，大きいキーワード集合から，K_gの検索要求がある場合に，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，検索に関係する全てのタプル集合 Tから，キーワード集合 $K = \{k_j | j \in J\}$ 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 Kは，検索に関係する全てのキーワード集合 Kの任意の要素の全ての組み合わせとなる．T_iにおける k_jの重みが w_{ij}である時，キーワード集合 Kのサポート値 $sup(K)$ は次式によって

求めることとする.

$$N_0 = \sum_{T} \max_{K} w_{ij},$$

$$N(K) = \sum_{i \in I} \min_{j \in J} w_{ij},$$

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

上述した手順によって，検索対象となるデータベースに与えられたキーワード集合から，相関のあるキーワード集合が導出される．なお，ここでは，以下の重みを用いた．

キーワード k_j で検索を実行した時，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 検索システムの実装

本章では，検索対象となる雑誌記事索引データの構成について示し，どのような手順で関連語・連想語検索が可能なデータベースを実装したかについて簡単に示す．

なお，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となっている．ここでは，それらのデータから，最近の 65,453 件のデータを実験に用いた．

4.1 雑誌記事索引データの変換

まず，国立国会図書館の雑誌記事索引データベースについて簡単に述べる．

- 記録媒体：磁気テープ
- 記録密度：9 トラック／6250BPI
- コード体系：1 バイトモード：EBCDIC, 2 バイトモード：JISX0208（1978 年版）
- ファイル形式：(ラベル：標準ラベル),(レコード形式：可変長ブロック),(ブロックサイズ：最大 12000 バイト),(レコードサイズ：最大 11996 バイト)

なお，データベースは，以下の属性をもつ．

（属性）コントロール番号，ステータス，更新日付，論題名，多言語論題名，著者名，著者名読み，雑誌名，多言語雑誌名，編者等，請求記号，雑誌巻号，刊行年月次，ページ，雑誌分類，雑誌種別，論題フリガナ，雑誌ＩＤ，ＩＳＳＮ，本文の言語

さらに，以上の属性値をもとに自動的に検索キーワードなどが作成されているが，今回は，データの冗長度が高いため幾つかの属性は用いなかった．

（属性）論題（漢）（フリ），著者名（漢）（フリ），著者名（英），雑誌名（漢）（フリ），他誌名（漢）（フリ），編者等（漢）（フリ），出版者（漢）（フリ）

次に，レコードサイズなどの制約を取り除いた上で，ワークステーション上で利用可能なフォーマットへと変換を行った．文字コードは EUC コードを選択し，データファイルはレコードごとに分解して HTML 形式により整形した．以下に，変換後の書式の一例を示す．

例：HTML 化されたデータの一部

```
<TITLE>
図書館情報学案内——
Carson,Paula P.et al.
 The library manager´s deskbook:102
 expert solutions to 101 common dil
emmas. American Library Association
(1995)</TITLE>

著者名:  <BR>
雑誌名: 国立国会図書館月報 <BR>
巻号: ４２８ <BR>
ページ: ３１～２９ <BR>
発行年: 199611 <BR>
```

本システムの実験により明らかになってきたが，タイトル部にギリシヤ文字が多用されるなど，今後，中国語，韓国語，ベトナム語，タイ語など多国語を用いた書籍情報の増加が考えられることから，多様な文字コード処理能力も必要と考えられる．

4.2 検索システムの構成概要

検索システムは，(1) データ収集・解析部，(2) データベース，(3)CGI (Common Gate Interface) を利用した検索要求処理サーバ，の各システムから構築される．

データ収集・解析部

ネットワークを通じて書誌データの収集を行うと共に，データの解析を行い，データベース (database) へと圧縮したフォーマットによって格納する．なお，収集された文書の解析では，HTML のタグによる文書の論理構造を用いることとした．ただし，構文解析や検索などに伴う計算コストを下げるために，ノイズを増加させると考えられる頻出単語や平仮名や数字をキーワードから除外し，解析されたデータをデータベースへと圧縮して格納することにしている．

検索処理サーバ

検索処理サーバ (query server) は，ユーザからの検索要求に対して応答する．まず，ブラウザ (browser) によって検索記述が要求されると，CGI(Common Gate Interface) を通じて検索処理サーバへ通知される．次に，データベースから条件を満すデータ集合を求め，それらのデータ集合に対してデータマイニング操作が行われる．なお，検索結果は，HTML 形式によってユーザへと提示される．

データベース構造

実時間で効率良く検索並びにキーワード導出が可能となるように，計算コスト面からも考察した上で，処理を実行するために必要となるデータベース・スキーマを示す．

・Keyword テーブル

$$K_m: \ (URL_{m1}, W_{m1}), \ldots, (URL_{mn}, W_{mn})$$

キーワード K_m に対して，ドキュメント URL_{mn} と重み W_{mn} の組のリストが格納される．

・URL テーブル

$$URL_m: \ Attr_m, (K_{m1}, W_{m1}), \ldots, (K_{mn}, W_{mn})$$

URL_m に対する，タイトルや作成日などからなる $Attr_m$ と，ドキュメントが含むキーワードと重みの組 (K_{mn}, W_{mn}) が格納される．

4.3 検索例

前節のデータベースから，タイトルに対する検索キーワードから相関ルールを導出し，本システムで採用した処理の妥当性の検証を試みる．なお，データベース中のキーワード総数は，54,904 語である．

図 3は，検索語として“環境”を与えた実行例である．本システムで検索された文献数は，1,911 件となっており，関連語として，

> 特集，地球，保全，運輸，問題，デザイン，白書，対策，監査

が提示されることとなり，絞り込み，または，関連語への移行の可能性を与えている．

なお，RCAAU を用いた場合，1,793,301 件の URL に関わるキーワードを解析して総キーワード数 393,774 語をもつ．図 4は，検索語として“環境”を与えた実行例であり，10,566 件の URL から，

> environment, 計算機，設定，動作，問題，開発，学習，公害，教育

などが関連性の高いキーワードとして導出され，上記と異なる語も含めて提示される．また，それ以外にも，電子ニュースなどの記事，メーリングリストの電子メールから求まる相関ルール [7] を，本検索システムに援用することによって，より良い図書・文献データベース検索の実現が行い得ると考えられる [8]．

このように，複数の検索システムを連携させることによって導出される相関ルールを用いることにより，検索式の可能性が広がるため，より好ましい全文検索が可能になると考えられる．よって，今後のナビゲーションシステムの構築にあたって，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をより的確に把握させ，入力キーワードが一般にもつ揺らぎを理解する機会を高めることが鍵となると考えられる．

5 結論と今後の課題

現在，電子図書館や電子出版などを始めとして，コンピュータ環境の変化と共に急速に電

子化が進められているが，文献情報に対する検索手段の提供は依然満足なものではない．特に，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る傾向にあるが，分類・主題分析などを行わない状況でのデータ格納は，まさに生データから知的な検索を行うデータマイニングに求められている技術そのものを必要とする．

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対して，領域知識が不足する場合に有効な検索の枠組みを与えることを目指しながら，データマイニング技術を基礎とした検索システムの構築を試みた．

今後，本稿で導出されたルールを検索ユーザへの知識として自然にフィードバックすることによって，より自然な検索過程を支援することが可能になる検索式精練アルゴリズムが重要な課題になると考えられる．

謝辞

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を御提供頂いたことに感謝する．また，日頃御指導頂く京都大学大学院工学研究科応用システム科学専攻 長谷川利治教授 に深謝致します．さらに，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，新須哲朗氏，土屋悟氏，花房寛氏に感謝する．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] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp.487–489, 1994.

[2] 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.

[3] 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.

[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), pp.91–97, 1997.

[8] 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," 情処研報 97-DBS-112, pp.33–40, 1997.

[9] 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.

[10] 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," Proc. of Advanced Database Symposium '96, pp.27–34, Tokyo, December 1996.

[11] 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.

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

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

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

[15] 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.

[16] 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.

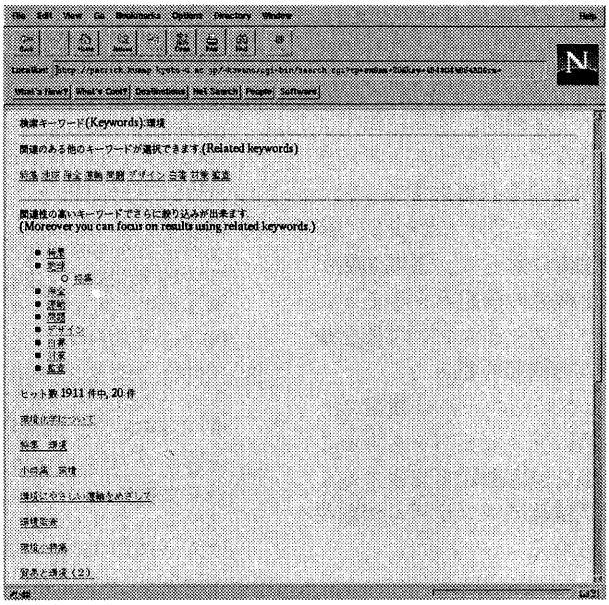

図 3: 本システムでの検索結果画面

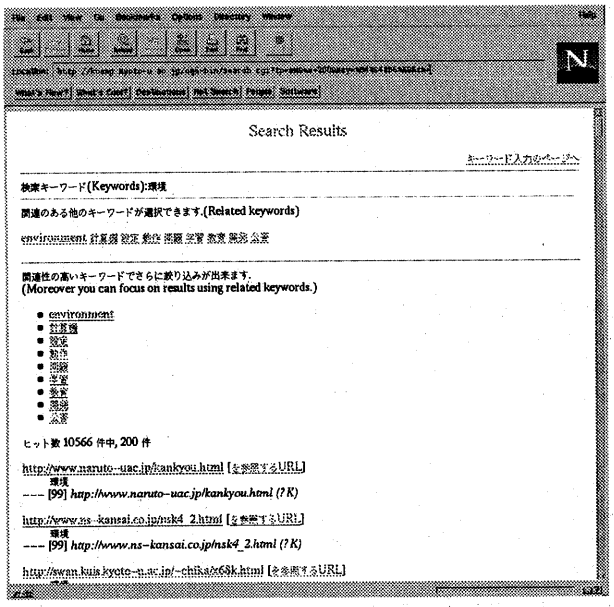

図 4: “問答” での検索結果画面