データベースシステム 113－40
（1997. 7. 15）

文献二次情報データベースにおける検索支援*

川原 稔† 河野 浩之‡

†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

文献二次情報データベースにおける文献検索は，検索に必要な領域知識や背景知識だけでなく，属性値の曖昧さのため一般的に難しい．そこで，本稿では，文献二次情報データベースの属性に対して意味付けを行い，それらに対して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方法について述べる．また，属性間の関連からのキーワード空間の拡大を行い，相関ルールを求めることにより，相関性が高いキーワード集合を得て利用する手法についても述べる．

キーワード：文献検索，データマイニング，相関ルール，属性

Implementation of Bibliographic Navigator with Association Rules

Minoru Kawahara† Hiroyuki Kawano‡

†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 a bibliographic database. One of the reason of the difficulty is that value of the attribute is not limited in certain values. In this paper, in order to provide more helpful knowledge, we categorize the attributes and extend the mining association algorithms which discover relevance to a keywords space. By our algorithms, interesting rules are derived from relationships between several categorized attributes in a database. We also propose algorithms in order to modify an initial query with keywords, which are specified by users' view. Moreover, we develop a navigator using textual data mining algorithms, and verify the effectiveness of our proposed algorithms.

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

* 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

図書や文献に関する情報をデータベース化した文献二次情報データベースにおいて，目的の文献を検索するのは，熟練した図書館司書に頼まないと見つからないなど，一般的には難しい．この困難さは，操作に関してシステムの特徴など背景知識が必要であることと，検索対象分野に関する領域知識が必要であることに起因している．そのため，困難さを解消あるいは緩和するための有効な検索手法に関する研究が，これまでにも多く行われてきた [8, 10].

しかし，計算機環境の発達により業務の多くが電子化され，文献情報あるいは文献そのものの電子化が急速に行われ，データが著しく増加する状況で，目的の文献を探し出すことはさらに困難になっている．そうした状況は文献検索に限らず，値の不確定な文書データを大量に処理する必要性が高まっており，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) [1, 5, 12] に関わる分野において，実用性の高いルールを精度良く導くことを目標に盛んに研究されている．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 [2, 5]. 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 [7] や，テキストデータ発掘 (textual data mining) の研究 [3] 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 [11] 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 [4, 5]. その手法を図書・文献データベース検索に適用する試みも行い，有効であることが確認された [6].

そこでは，文献登録システムを対象とした，文献の内容に関する情報を多くはもたないデータベースから相関ルールを求めるため，異種データベースを用いてキーワード空間を補完した．本稿では，属性の付与や文献に関する情報ばかりでなく，キーワードやアブストラクトなどの内容に関する情報が付加される文献二次情報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，属性から相関ルールを有効に取り出すために，含む情報により属性を特徴付ける．そして，各属性から導出される相関ルールと属性の特徴付けを基に，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．

以下，2 章では，文献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，文献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．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 文献検索システムの問題点

文献二次情報データベースでは，通常のデータベースと異なり，基本的に文書データであるため属性値の値域が制限されない．また，著者や出版社により属性値の与え方が異なり曖昧となる上に，データベース編纂者の分類方法によって属性や属性値が異なっている．そのため，文献検索ユーザは，用いるべき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．[8].

そこで，より優れた文献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のは難しく，抽出データにゆらぎが発生してしまう [8]. また，文書ベクトル空間に対する処理を行う検索 [9, 10] の検索評価基準として，再現率・適合率を用いることもあるが，その手法は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必要とするアルゴリズムが多く，大規模なデータに対する適用は現実的でない．

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

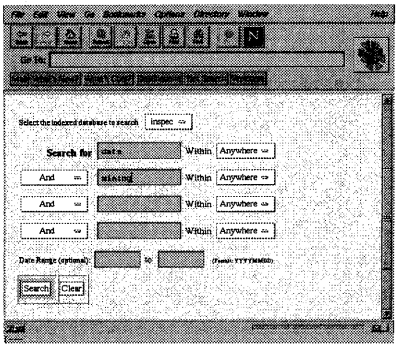

図 1: 現在の文献検索システムの例

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として，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．しかし，組織や著者により単語に対する位置付けが異なるため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り，それら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．

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，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，シソーラス展開のような一般性の高いキーワード空間から得られる相関ルールも利用することも考慮して，実験システムに INSPEC データを用いながら，文献二次情報データベースに対して，より良い検索式の改善を行う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 複数属性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文献二次情報データベース検索支援システムでは，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，

表 1: 属性の分類

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，データベース編集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:出版社，会議名，ISBN

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．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 [6].

3.1 文献二次情報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，文献情報データベースにおける属性の特徴を整理し，複数の属性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．本稿では，属性を表 1 のように 3 つのタイプに分類し，このうち特徴抽出属性と内容記述属性を用いる．

タイトルなどの特徴抽出属性には，文献に対する高い相関性をもつキーワードが含まれているが，属性値に含まれる単語数が少なすぎるため，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，関連する属性を収集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 [6]．また，相関ルールによる関連キーワード導出では，多くが無意味語となりがちであり，その除去も考慮しておく必要がある [6].

3.2 複数属性を用いた検索式生成アルゴリズム

特徴抽出属性は，著者あるいはデータベース編纂者により，その文献の特徴を示すキーワードが与えられるため，得られる相関ルールは高い相関性を示すキーワード集合 $\mathbf{K_f}$ を与える．

次に，予め指定された関連する属性を収集して生成されるキーワード空間から相関ルールを導出し，キーワード集合 $\mathbf{K_a}$ を導出する [6]．特徴抽出属性を基にして得られるキーワード集合 $\mathbf{K_a}$ 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．さらに，著者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．

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 $\mathbf{K_g}$ 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 $\mathbf{K_g}$ 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，より一般的に成立するルールと捕らえることができる．

さて，これらの相関ルールから求まるキーワード集合 $\mathbf{K_f}$，$\mathbf{K_a}$，$\mathbf{K_g}$ は，丁度，異種データベース利用における文献データベースおよびローカルデータベース，グローバルデータベースにそれぞれ相当する [6]．したがって，これらのキーワード集合から，同様にして関連キーワードを抽出することができる．

4 相関ルールを用いた検索支援システムの構成

本章では，文献二次情報データベースとして INSPEC データベースを用いて，検索支援システムを構成する．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，本実験システムには最近の約 3 年分の 968,230 件のデータを格納している．

この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．

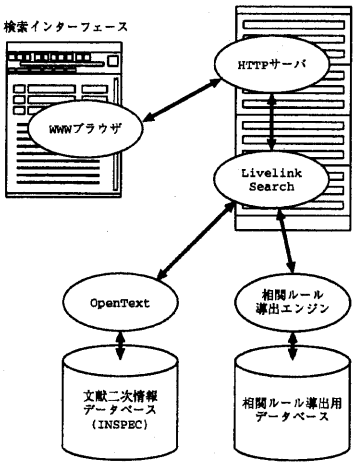

図 2: 本システムの構成

表 2: 属性とキーワード集合の対応

属性	タグ	キーワード集合
タイトル	Title	$\mathbf{K_{in}}$
キーワード	Keyword	$\mathbf{K_f}$
著者 (関連)	Title ⇒ Author ⇒ Title	$\mathbf{K_a}$
アブストラクト	Abstract	$\mathbf{K_g}$

INSPEC データベースに対する文献検索は，WWW ブラウザから OpenText への HTTP インターフェースである Livelink Search を介して OpenText に対して問い合わせを発行することで行う．Livelink Search に含まれる CGI において，OpenText に対して問い合わ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 (図 2)．

検索式の改善方法が，絞り込み，あるいは，関連語への移行の形でブラウザ上に提示されるので，検索ユーザは，ほとんど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，簡単に改善された検索式による検索を行うことができる (図 4)．

相関ルール導出は，表2に示す属性を用いて行った．このうち，キーワード集合$\mathbf{K_a}$は次の手順で求める．

1. 入力検索式に含まれるキーワードをタイトル属性値に含むタプル集合を求める．
2. 各タプルについて，そこに含まれる著者を著者属性値に含むタプルを求めて，タイトル属性値に含まれる全キーワードからキーワード集合を生成する．
3. 各タプルごとに生成されたキーワード集合からキーワード空間を生成し，それに対して相関導出ルールを適用して，関連のキーワード集合を導出する．

相関ルール導出用には，予め表2に示した各属性に対して属性値のキーワード解析を行い，キーワードとその重みに関するデータベースを構成した．検索時には，そのデータベースを用いて相関ルールを導出する．

図3と図4は，検索語として“bibliographic”を与えた実行例である．解析内容を見ると，各キーワード集合は表3のようになっており，その結果，関連語として，

database, information, system, data,
library, online, record, retrieval

が得られた．この結果がブラウザ上に提示されて，元の検索式に対して，絞り込み，または，関連語への移行の可能性を与えている．提示された関連語と表3を見比べると，タイトルのみから導出されるキーワード集合$\mathbf{K_{in}}$に対して，“library” および “data” が$\mathbf{K_f}$により与えられていることが分かり，キーワード空間の狭いタイトルだけでは得られなかった関連語が抽出されたことが分かる．

表 3: キーワード集合の結果内容

キーワード集合	内容
$\mathbf{K_{in}}$	information, database, record, system, online, retrieval
$\mathbf{K_f}$	information, library, database, system, online, data
$\mathbf{K_a}$	database, information, system, retrieval, online, based
$\mathbf{K_g}$	information, library, system, database, data

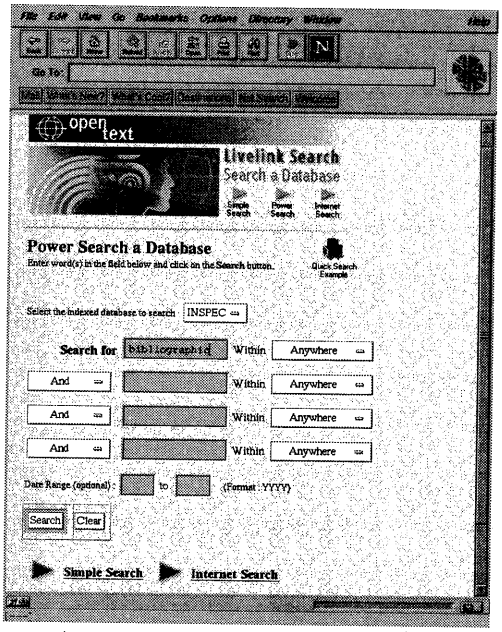

図 3: 本システムでの検索画面

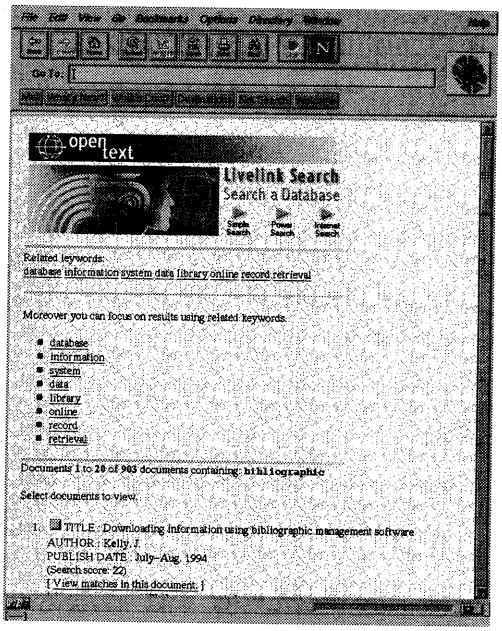

図 4: 本システムでの検索結果画面

本実験では，属性の位置付けを考慮した相関ルール導出により，検索式が改善され，より好ましい全文検索が可能であった．また，その結果から，入力キーワードがもつ揺らぎを把握する機会を得ることができた．

5 結論と今後の課題

電子図書館や電子出版などが注目を集める昨今ではあるが，増加しつつある大量の文献情報から目的の文献を取り出すことは，対象とする文献が膨大になる分さらに難しくなっきているといえる．その増大する情報に追随できなくなり，文献情報に関して，これまでのような分類・主題分析やキーワード統一も行わないフリーキーワード検索が増えており，状況はさらに悪化している．

そこで，本稿では，文献二次情報データベースに属性値として与えられる情報を利用して，データマイニング技術を基礎に領域知識を補うことにより，有効な検索が実現できるシステムの設計指針を示すことができた．また，シソーラス展開など一般性をもたせた付加データの利用にも考慮した．アルゴリズムの計算量を十分に抑制しており，現在稼動する計算機システムで実用的なレスポンスタイムで実現する検索システムの構築が可能である．

今後，RCAAU システムの機能連携システム [6] の機能との有効な使い分けを行い，検索対象などに応じた動的な重み付けによる，より有効な検索結果の導出を行うアルゴリズムを洗練する必要がある．

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．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 に深謝する．

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

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

[3] 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.

[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 回データ工学ワークショップ (DEWS'97), 1997.

[5] 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.

[6] 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.

[7] 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.

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

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

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

[11] 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.

[12] 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.