データベースシステム 108－5
（1996. 5. 23）

WWWデータ資源検索におけるデータマイニング手法*

河野 浩之† 長谷川 利治†

†京都大学大学院工学研究科応用システム科学専攻

WWWシステムには，複数の視点による緩やかなリンクをもつ膨大なデータが蓄積されており，必要となるデータをスムーズに検索するために，多くの背景知識が要求される．我々は，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，検索要求キーワードから導出されるルールを検索支援知識として与えた．本稿では，検索システムでの実験をもとに，WWW検索において属性指向アルゴリズムや組指向アルゴリズムを用いる各種データマイニング手法がどのような意味をもつかについて論じる．また，ネットワーク環境を考慮した分散型検索システムを構成する場合に，データマイニングのもつ意味を，ネットワークの品質に関する問題に注意を払いながら述べる．

キーワード:WWWデータ検索，データマイニング，分散環境，重み付き相関ルール，データベースからの知識発見

Data mining technology for WWW resource retrieval

Hiroyuki KAWANO† Toshiharu HASEGAWA†

†Department of Applied Systems Science, Faculty of Engineering, Kyoto University

Without rich background knowledge, it is very hard to discover useful WWW resources which are loosely linked in the network. With applying techniques of data mining, such as weighted association algorithm, to the WWW resource retrieval, it is possible to derive useful rules in order to modify users' search queries. In this paper, for constructing distributed WWW search systems, we consider the quality of search results in the point of data mining methods, such as attribute-oriented algorithms, tuple oriented algorithms and others. This paper also includes a brief discussion of mining rules regarding quality of communication channel.

Keywords: WWW search engine, data mining, distributed environment, weighted association rule, knowledge discovery in databases

*連絡先: 〒606 京都市左京区吉田本町 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

現在，コンピュータネットワーク上で膨大なデータ資源が緩やかなリンクによって相互に結合されながら提供されている．特に，WWW (World Wide Web) は，マルチメディアデータを用いた自由度の高い情報発信が可能であるため，そのサイトの増加は目覚ましい．これまでに種々のWWW検索システムが提供されているが，精度の高い検索を実行するためには，かなり質の高い検索記述が要求される．また，提供されるWWWデータの自由度が高いため，複数組織のサイトに分散してデータがある場合には，データ相互の関連を把握することは非常に困難となる．

そこで，これまでに研究を行っているデータベースからの知識発見 (KDD: Knowledge Discovery in Databases) 技術やデータマイニング (Data mining) 技術 [3, 1, 6] を用いて，WWWデータからルールなどを導出し，検索ユーザに対してWWWデータの関連を提示したり，検索記述の支援を行うなど，WWWデータ検索支援を対象にした実験を行っている．ロボット [4] と呼ばれるプログラムによって，WWWデータと，そのリンク構造を収集し，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 [9] を拡張することによって，ルールを求める．得られる相関ルール (association rule) は，検索ユーザが用いたキーワード集合を含むWWWデータとの関連性が強いキーワード集合を与えるものとなっている．得られたルールによって導かれるキーワード集合を，検索ユーザに対して提示することによって，連想的検索を実現している．本システムに関するURLは，“http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/” である．

そこで，本稿では，WWWデータの検索結果として，データマイニング技術を用いたデータとルールを提示する際に，どのような点を考慮すべきかについて議論する．また，データマイニング手法は計算コストが非常に大きいことが実験システムの実装において問題であったため，対象となる検索集合の構造を簡約化するために，どのような背景知識を用いるべきかについて述べる．さらに，分散環境において検索システムを構築する場合に，ネットワークの品質を，どの様に考慮しなければならないのかについても示す．

以下，2章では，WWW検索システムの問題点を述べ，我々が検索システムを構築する上で，どのような方針を採用したかについて触れながら，実験システムの構成と特徴を簡単に述べる．3章では，重み付き相関ルール導出アルゴリズムを含めてデータマイニングアルゴリズムのもつ意味を考え，検索結果としてどの様なデータとルールが検索ユーザに対して提示されるべきかについて述べる．4章では，分散協調型の検索システムを構築するために問題となる点を，ネットワーク環境を考慮したデータマイニング手法の観点を含めて整理する．

2 WWW検索システム

現在，データベースシステムと連動させながらWWWデータが提供されつつあり，そのため，単一の組織内で提供されているデータに関する検索は，次第に容易になりつつある．しかしながら，WWWデータが高い自由度で提供できる特性を考えた場合，広域WWW検索システムである，Alta Vista, InfoSeek, Lycos, Harvest, WebCrawler, RBSE spider, Fish-searchなどの多くのシステムは非常に重要な役割を果たしていくと考えられる．さらに，WWWデータが膨大になるに従って，既存の多くのシステムが提供している単純な検索サービスだけではなく，複数組織やサイトに存在する多くのデータの質を総合的に評価し，多面的な分析を加えた上で，それぞれのデータの位置づけを検索者が把握できるような検索支援サービスが必要となっていくと考えられる．

2.1 WWW検索における問題点

WWWデータの急速な増大は，検索対象に関する知識を十分にもたないユーザが検索シ

ステムを利用する際に，目標となるデータへと達する検索の遂行を困難としている．この問題は，既存の検索システムに共通する以下の項目に起因するものと考えられる．

1. 目標に達するまでに必要となる問合せ記述能力やキーワードなどに関して，ユーザ自身に強く依存しており，検索支援環境が提供されていない．
2. ハイパーテキストにおけるリンク構造が，WWW データの収集以外に十分利用されていない．
3. 検索結果の表示は HTML(Hyper Text Markup Language) 記述であるため平面的であり，多量の検索結果を把握するための GUI が無い．
4. ブラウザが動作しているクライアント側の計算機資源が，殆んど有効に利用されていない．

そこで，データマイニング手法を用いることによって，検索記述や用いられている検索キーワードなどに関する知識をユーザへとフィードバックし，対象に関する問合せ記述の質を改善する検索支援システムの開発を行っている．既に，文献 [5] では，上述の項目 1. 及び項目 2. に焦点をあて，検索ユーザが与えたキーワードに対して関連性の深いキーワードを導出して，検索式を改善する支援について論じた．さらに，文献 [2] において，項目 3. 及び項目 4. について考察し，クライアント側で Java 言語の applet を利用した可視化 (visualization) を行うことにより，検索結果として提示されるデータ集合の特徴把握を容易にし，クライアント側での分散処理を実現したことについて述べた．

2.2 WWW 検索システムの構成

従来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てデータモデル設計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可能であった．しかし，WWW データでは，管理者によって統一されたモデルに従って提供されたものではなく，提供されている WWW データ自体のばらつきが大きいため，良い検索式を与えることが非常に難しくなりがちである．

さらに，全てのデータが偏り無く提供されている訳ではなく，また，実際に提供されているデータの分布状態に関する知識を得ることも難しい場合が多い．例えば，実世界では出現頻度の低い用語であっても，WWW データで頻度が高いキーワードで検索を行った場合に，大きな被覆をもつ検索結果として大量のデータを得ることになってしまう．つまり，実世界の「常識」と WWW データ中の「常識」とが異なっているため，データの偏りまでを考慮した精度の高い検索式を記述することは非常に困難であると言える．そこで，効率良く有用度の高いデータを検索するためには，質の高いキーワードを知ることが非常に重要となる．

我々は，これらの問題点の解決を目指した実験システムを構築しており，その構成について図 1を用いて簡単に述べる．各エージェント (agent) は，サーバ単位の幅優先探索により，被参照回数が多く有用度が高いデータから優先して取得し，日本語形態素解析システム JUMAN を利用したパーサ（parser）によって，「URL，キーワード，重み」などからなるデータをデータベースへと格納する．一方，URL にコード化された検索キーワードからなる問合せ記述を受け取った検索システム (query server) は，その検索結果を HTML 形式に整形してユーザーへと送信する．

なお，現在の実装では，利用者は正 (positive) のキーワード集合 K_p，負 (negative) のキーワード集合 K_n によって，AND, NOT 条件を用いた問合せ記述による検索を実現している．OR 検索は，同様の概念をもつ単語では先頭部分に語根が多いことと，複数形への対応を考慮し，前方部分が一致するキーワードに対して可能としている．これは，検索集合が非常に大きくなる場合や，検索対象に関係する適切なキーワードに対する知識が不足している場合に，OR 検索条件を満す検索式記述が困難となることを考

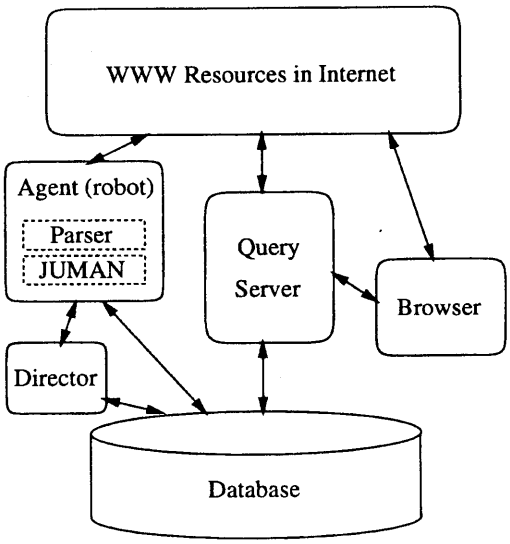

図 1: 検索システムの構成

慮したものである.

そして, 検索対象や関連するキーワード集合に対して質の高い背景知識をもたない場合にも検索が容易に遂行できるように, 重み付き相関ルールによって導出したキーワード集合をユーザに提示することによって, AND 条件の検索式記述の支援を行っている.

また, 実験システムにおいて現在収集中の検索結果からは, NOT 条件記述に適切なキーワードを指定することも難しいことが伺われるが, この点を含めてシステムから得られているデータの分析に関しては改めて報告を行う予定にしている.

さらに, 異なる管理組織に存在する殆んどのハイパーテキストにおけるリンクは一方向性であるため, 実験システムでは, 検索された URL を参照するリンクをもつ URL 集合の検索を可能としている. この様なリンクを逆にたどる検索機能を, 通常の検索結果と同時に利用することによって, より有用度の高いリンクへのアクセスを発見できる.

以下, 実験システムの特徴を簡単にまとめる.

1. 検索支援を行うルールによって, 関連語をリアルタイムに導出することを実現した.

2. ルールとして与える関連語は, 検索結果としてユーザ側に送信されるデータ以外のデータを用いて導出しており, 送信されない部分に関する知識を与えるものでもある.

3. 検索結果は GUI を用いて表示され, 検索ユーザ側にあるブラウザ内での検索を可能とした.

3 データマイニング手法

前章では, 重み付き相関ルールを求めるアルゴリズムを用いた WWW 検索支援システムについて述べた. 本章では, より一般的にデータマイニングアルゴリズムのもつ意味に関して述べるとともに, 文書検索との関連についても簡単に触れる.

3.1 データマイニング・アルゴリズム

データマイニングに関する研究が, データベース特有の性質を考慮して行われており, 組指向アルゴリズム [9] や, 背景知識として概念階層を用いる属性指向アルゴリズム (attribute-oriented induction algorithm)[10] などに関する研究が盛んに行われている [8]. そこで, 我々も知識発見アルゴリズムに関する研究を行っており, 多くの研究がなされている組指向アルゴリズムの一つである, 相関ルール導出アルゴリズム [9] を拡張し, 重みをもつキーワード集合からルール導出を行っている [5].

3.2 WWW データに対するデータマイニング

一般的に, 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. この場合に, 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, データマイニング技術を用いて, 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が重要である．この時与えられるルール集合 $\{r\}$ は，各属性間に成立する制約 (constraint)，規則性 (regularity) などを求めることによって与えられる．

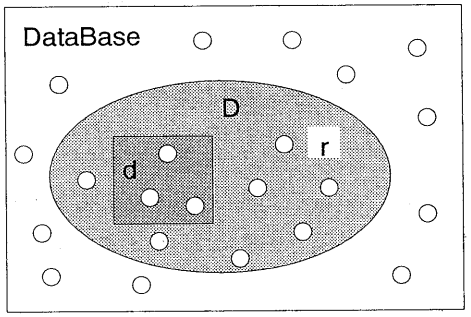

図 2: 検索結果と表示されるデータの関係

ここで，各々のデータのもつ情報量 [7] を $I()$ によって評価するならば，$I(\{D\}) \simeq I(\{d\},\{r\})$ を満すルール集合が与えられることが望ましいと言える．しかしながら，ノイズや例外を含まない理想的なデータであっても，一般に等号を成立させることは計算量の関係から非常に困難であるため，現実的には検索集合自身のもつ性質を考慮しながら，より質の良いルール集合を与えることが非常に重要となる．

被覆性を考慮したデータ生成

WWW データのように統一された管理がなされておらず，キーワードのばらつきが大きくなる対象については，検索ユーザの与えたキーワード集合に関係するルール導出を行って，検索式の精度を向上させるルールを与える手法が有効であると考えられる．つまり，図 3に示したように，与えられたキーワードに対する検索結果である $\{D\}$ 中で，多くのデータを被覆するキーワード集合 $\{\mathrm{k}_i\}$ をルールとして求めて，検索結果 $\{D\}$ の部分集合であるデータ集合 $\{d\}$ と共に，検索ユーザに与えることとなる．この時，$\{D\}$ の中で，$\{k_i\}$ を用いて被覆されない部分を，集合 $\{d\}$ として検索ユーザに提示すれば，マイニングによる損失は少なくなる．

再現率・適合率の利用

文書検索においては，検索結果として求められる単独の文章の質を重要視しており，再現率・適合率を検索評価基準において，各々の文書のもつ次元の高いベクトル空間を用いた検索を実現している．よって，ユーザに対する検索結果としてデータ集合 $\{D\}$ の部分集合である $\{d\}$ を検索結果として与える場合に，既存の文書検索の方法が有効であることも多いと考えられる．

なお，WWW データも，文書検索の対象となる文書の一種であるので，抽出された重要な単語と，それらの単語の出現頻度などを重みとして文書ベクトルを構成するならば，我々の提案したアルゴリズムを用いたルール導出が可能である．しかしながら，データマイニングでは，個々の文書の質を評価するのではなく，検索対象となった全文書の集合内に成立しているルールを求めることを目標としている点が大きく異なることには注意しなければならない．

概念階層・シソーラス展開の利用

さらに，この種のデータマイニングアルゴリズムにおいて，被覆するデータ集合の特性をより一般的に把握するために，属性指向アルゴリズムの概念階層と同様に分類 (taxonomy) を行う背景知識を用いるアルゴリズムが提案されている [9]．ただし，このような背景知識は，文書検索におけるシソーラス展開と同様の処理を与えるものであり，展開されるキーワードの質によって検索結果が大きく変化することに注意すべきである．

従って，概念階層などの背景知識 $\mathcal{K}$によって与えられる情報量を考慮すると，$I(\{D\}) \simeq I(\{d\},\{r\},\mathcal{K})$ を満す集合が検索結果として与えられることが望ましい．また，検索サーバ側の知識 $\mathcal{K}_s$と，クライアント側の知識である $\mathcal{K}_c$が常に等しいとはかぎらなので，一般には，$I(\{d\},\{r\},\mathcal{K}_s) \geq I(\{d\},\{r\},\mathcal{K}_c)$ となると考えられる．

ハイパーテキストのリンク構造の利用

ハイパーテキストの特徴であるリンク構造を利用したルールを導出する手法も考えられる．

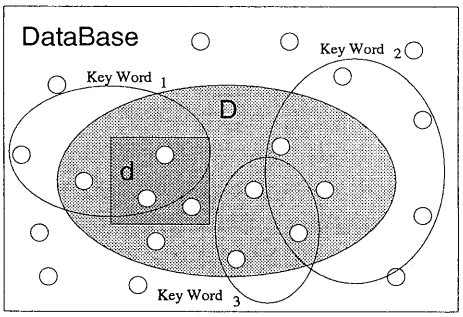

図 3: キーワードによる被覆を考慮したルール生成

このようなリンク構造を背景知識$\mathcal{K}$として利用するためには，同一組織や異なる組織におけるドキュメント間のリンクによる到達性や，到達に必要なステップ数を制約として用いることとなる．例えば，図 4中の黒丸で表したドキュメント集合を，検索結果のデータ集合$\{d\}$として与えることが可能である．リンク構造を利用した１ステップ到達性が保証されていることを，背景知識$\mathcal{K}$として共有していれば，データ集合$\{d\}$にアクセスした後，１度リンクをたどれば十分であることが分かる．

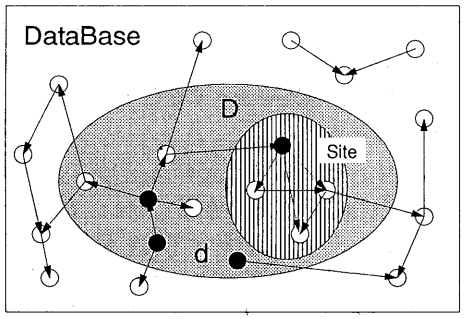

図 4: リンク情報を利用したルール生成

以上，データマイニングを行う上で，質の良いルール集合$\{r\}$を検索ユーザへと適切に与えることが最も重要である．そして，このルール集合と，検索キーワード集合に良く一致する集合$\{d_k\}$や，データ集合$\{D\}$への到達性の保証されたデータ集合$\{d_r\}$を適宜提示すれば，検索ユーザが効率的に検索結果を把握できる．また，質の高い背景知識$\mathcal{K}$を用いることによって，検索結果$\{D\}$に対してマイニング処理を行う際の計算コストを減らすことも必要である．

4 分散環境におけるデータマイニング

前章で述べたデータマイニングアルゴリズムを分散環境で実行するには，データ収集・検索を分散させた場合のネットワーク負荷のコスト評価に加えて，検索サーバiが異なる背景知識をもつ場合などについても考慮しなければならない．つまり，データベース$\mathcal{D}$において，あるキーワード集合によってデータ集合$\{D\}$が検索結果として求まった場合，ユーザに対する検索結果としては，どの程度のデータ転送が可能であるかを評価し，適切にデータ集合$\{d_i\}$とルール集合$\{r_i\}$を絞り込む必要がある．加えて，どの様な背景知識$\mathcal{K}_i$を利用することによって絞り込みが行われたのかについても評価する必要が生じる．

4.1 ネットワーク環境を考慮した検索

現在の WWW 検索システムは集中型の処理による検索支援を提供しており，図 5に示したようなネットワークにおいて，全てのクライアントがサーバへの問合せを実行していることとなる．

なお，集中型の検索サーバを利用する場合においても，ネットワークにおけるクライアント側の位置によって，アクセス対象となる WWW サーバへの接続コストが異なっていることから，検索されるべき結果は相対的なコストを含めて表示するべきである．そこで，我々は，通信ネットワークの状態を反映した検索環境を提供するためのブラウジング機能を実装し，大量に検索される URL に対して，クライアント側のネットワーク位置に依存したアクセスコストを含め

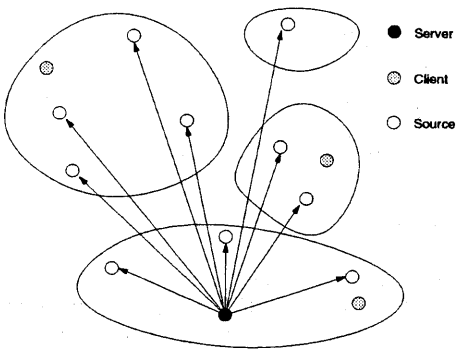

図 5: 集中型 WWW 検索システム

た検索を可能とした [2].

さらに，分散環境下においてより効率的な検索を実行するためには，通信コストなどを含めた種々のコストを的確に反映させた評価が重要となる．そこで，通信ネットワークの伝送速度に応じた処理に関して考えるならば，高速な送信が可能な場合には，クライアント側への大量データの伝送も可能と考えられる．この時，サーバ側の背景知識 $\mathcal{K}_s$の質によっては，ルールが劣化すると考えられることから，クライアント側におけるマイニングアルゴリズムの実行を重視することも考慮すべきである．一方，低速回線を利用する場合には，サーバ側で多くの処理を行った後で，クライアント側に送信しなければならない．

しかしながら，クライアント側の計算機資源を，さらに有効に活用するためには，データの転送コストのみならず，次の点に対する考察が必要となる．(1) 問い合わせに対する検索結果をクライアント側に保持することによって，通信コストを低減する．さらに，(2) 検索結果に対してマイニングアルゴリズムをローカルに実行することによって，サーバ側の計算コストを低減する．加えて，(3) マイニングによって新たに必要となるデータの転送要求を行う高度な問合せ記述を生成する．

以上のように，ネットワーク環境やマイニングアルゴリズムの計算コストに応じて，サーバとクライアント間で高度な動的負荷分散を実現することが必要である．

4.2 分散環境におけるデータマイニング

本節では，複数の検索サーバがネットワーク上に分散して配置される場合について述べる．

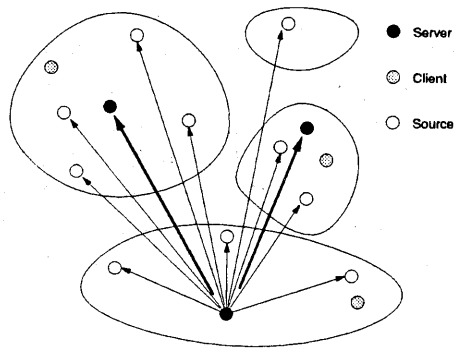

図 6: 分散型サーバによる検索 (I)

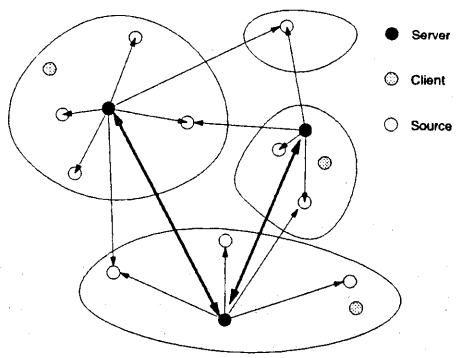

図 7: 分散型サーバによる検索 (II)

まず，遅延時間の特性を考慮した場合，遅延時間の平均·分散が大きい場合には，ミラーサーバを設けるなどして実時間での検索を可能とする必要がある．この時，図 6に示したように，複数のサーバに同一の内容をミラーする構成をとり，クライアントは，もっとも近いサーバに

アクセスを行う．一方，遅延時間の平均・分散が小さく安定している場合には，図7に示した複数サーバが協調動作を行う分散検索システムを構成することが可能である．

さらに，図7において，複数のサーバがネットワーク内の限定された領域のデータ収集を行う場合を考える．まず，複数サーバにおいて同一のアルゴリズムや背景知識を用いて処理する場合は，サーバ間のデータ交換を協調するだけであり，ミラーサーバを構成するのと同等である．一方，複数のサーバが，異なるアルゴリズムや背景知識を用いることも可能である．しかし，背景知識を共有することが困難である場合は，データマイニングによって導出されるルールの質の劣化は大きくなりがちである．この時，クライアントは，検索結果として得られるルールの質を評価しながら，サーバの組み合わせを動的に変化させて検索する必要が生じる．

今後，本章で述べたように，ネットワークと計算機資源のコストに応じて適切な比率でデータマイニング処理を分散させて，

$$I(\{D\}) \simeq \sum_{i=1}^{\#ofservers} I(\{d_i\},\{r_i\},\mathcal{K}_i)$$

の制約を満す，データマイニング手法の研究が必要である．

5 おわりに

本稿では，データマイニングアルゴリズムの意味について，検索ユーザに提示するデータ，ルール，そして，背景知識からなる $(\{d\},\{r\},\mathcal{K})$ に関する考察をおこなった．なお，現在の実験システムは，頻度の高い単語を除去したり，曖昧性の高い略語を除去するための背景知識 $\mathcal{K}_s$ を用いており，ユーザの検索キーワードと一致する $\{d\}$ と，重み付き相関ルールによる $\{r\}$ を与えている．今後，分散環境において効率良くデータマイニングを行うために，広域ネットワークにおけるWWW検索システムを通信コストに関する評価問題の一例として取り上げながら，分散協調型のデータマイニング手法に関して研究を進める必要がある．

謝辞

本稿の基礎となる実験システムの開発を行った，京都大学大学院応用システム科学専攻の西村英樹氏（現在，シャープ株式会社）と伊藤耕一郎氏に感謝する．

参考文献

[1] Holsheimer, M. and Siebes, A., "Data Mining – The Search for Knowledge in Databases," CWI Technical Report CS-R9406, 1994.

[2] 伊藤耕一郎，西村英樹，河野浩之，長谷川利治，"重み付き相関ルール導出アルゴリズムをもつ検索インタフェースのWWWデータへの適用，" 電子情報通信学会 1996 年総合大会，Vol.D, pp.307-308, 1996.

[3] 河野 浩之，西尾 章治郎，Han, J.，"データベースからの知識獲得技術，" 人工知能学会誌，Vol.10, No.1, pp.38–44, 1995.

[4] Koster, M., "Guidelines for Robot Writers, " http://info.webcrawler.com/mak/projects/robots/guidelines.html.

[5] 西村英樹，伊藤耕一郎，河野浩之，長谷川利治，"重み付き相関ルール導出アルゴリズムによるWWWデータ資源の発見，" 第7回データ工学ワークショップ (DEWS'96), 1996.

[6] 西尾章治郎，"大規模データベースにおける知識獲得，" 情報処理，Vol.34, No.3, pp.343-350, 1993.

[7] 小田島 潤，河野 浩之，長谷川利治，情報理論的な探索基準をもつ属性指向アルゴリズム，第6回データ工学ワークショップ (DEWS'95), pp.151–158, (1995).

[8] Piatetsky-Shapiro, G. and Frawley, W. J., "Knowledge Discovery in Databases," AAAI/MIT Press, 1991.

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

[10] Zaiane, 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), Montreal, Canada, pp.331–336, 1995.