多次元的なログデータマイニングを実現するデータキューブ機構の提案と評価

成瀬　正英† 大森　匡† 星　守†

† 電気通信大学大学院情報システム学研究科　〒 182–8585 東京都調布市調布ケ丘 1–5–1
E-mail: †{naruse,omori}@hol.is.uec.ac.jp

あらまし　近年，稼働し続けるシステムや Web サーバから生成される膨大なログを分析し理解する必要性は高まってきている．我々の提案するアイテムセットキューブ機構では，ログ内に記録されている事象から，多次元的な分析条件に対応した出現頻度の高い事象の組合せの集合を高速に算出する．この組合せを用いることで，数値による分析だけでは解明できない Web サイト上でのユーザの振舞やシステムの異常を知ることができる．また，この機構には多様な問い合わせに対し，キューブを変形させて高速に応答する OLAP 的な特性が備わっている．本稿では，この機構の実データへの適用とその評価を行った．

キーワード　Itemset Cube, データマイニング，OLAP, ログ分析

A New Data Cube System for Multi-Dimensional Log Data Mining

Masahide NARUSE†, Tadashi OHMORI†, and Mamoru HOSHI†

† The University of Electro-Communications, GraduateSchool of Information Systems　Chofugaoka 1–5–1, Chofu, Tokyo, 182–8585 Japan
E-mail: †{naruse,omori}@hol.is.uec.ac.jp

Abstract　Recently there is much need of understanding logs which are generated by computer systems or web servers. To solve this need, this paper proposes a new datacube system named Itemset Cube. It is used to understand logs by data mining through a multi-dimensional analysis technique like OLAP. Definition and algorithms of the itemset cube and its applications on a real dataset are described.

Key words　Itemset Cube, data mining, OLAP, log data analysis

1. はじめに

今日稼働し続ける計算システムや Web サーバは大量のログ (記録データ) を生成しており、そのログ分析を通してシステム内で起きている現象を理解したいという要求は高まっている。一般に、ログデータは、どのような種類の事象の組み合わせが発生したのか、発生の時刻、どのマシンやどのユーザから生成されたか、などの多属性から構成されている。そのため、最も多いログ分析法は、注目する属性の組を決め、それ以外の属性を無視して、何らかの数値指標について適切な集計処理を経てグラフ表示することである [5] [10] [8]。例えば、Web サーバのアクセスログ分析では、ユーザの訪問 1 回の記録を表すセッションレコードを対象に分析を行うが、このレコードは、アクセスしてきたユーザのインターネットドメイン、アクセス開始時刻、アクセスした Web ページの集合 (または列)、などの属性を持っている。典型的なグラフ表示では、ページをいくつかの種類に大別しておき各ページ種類あたりのアクセス数を時間軸に沿って表示したり、あるいは、ユーザドメイン別に Web サイトへのアクセス数を時間軸に沿って表示する、などがある [8]。この手法は単純だが、ログの属性数が増えた時に多様な属性の組み合わせについての集計問い合わせを即座に行うことは容易ではなく、データキューブ (OLAP) [14] が必要になる.

一方、他のログ分析法として、何らかの基準で決めた着目状況において生じている現象を、データマイニング技法の 1 つである高頻度アイテムセット分析により調べる手法もある [6] [12] [13] [7]。Web アクセスログの例では、「申込みページをクリックしたユーザが他によく見たページ集合を上位 10 個を求めよ」、システムエラー解析なら「サーバダウン前 10 分間に生じたエラーメッセージの頻出組合わせ上位 10 個を出力せよ」などである。

多属性を持つログデータ集合を理解するためには、上記 2 つの分析処理を、様々な属性の組み合わせや属性の分割方法 (時間を週単位で分けて見るか日単位にするか、など) の下で繰りかえし行う必要がある。

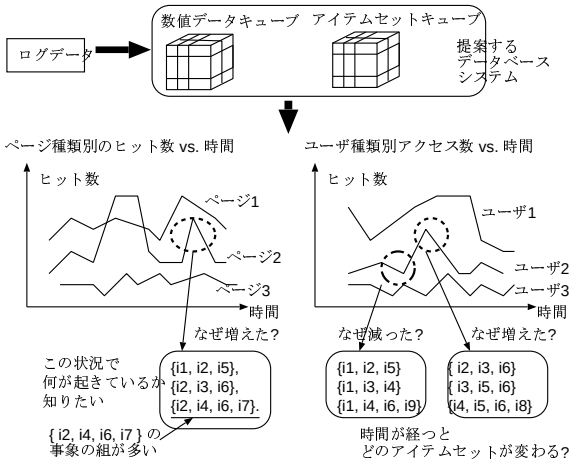


図 1　アイテムセットキューブシステムの目的

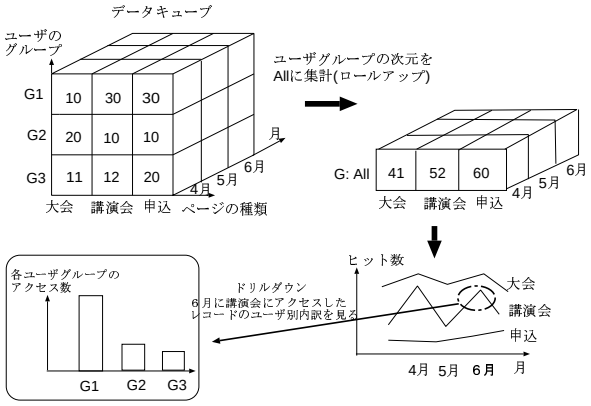


図 2　数値データキューブによる分析

そこで、著者らは、上記の要求に答えうる汎用的なデータベースシステムとして、アイテムセットキューブと呼ぶデータキューブシステムを提案している [1] [3]。その目的は次の 2 点である：

*1. まず、従来の数値データキューブによって、多次元ログデータから多様な属性の組について、集計問い合わせを行って様々な数値グラフを作成することを利用者に許す。

*2. 次に、利用者は、適当な数値グラフにおいて着目すべき状況を任意に決める。そして、その状況で生じている主要な現象を、即、アイテムセット分析により計算する。

図 1 は、この目的を満たすデータマイニングサーバの利用方法を、Web アクセスログ分析の場合について描いたものである。図中、利用者は、数値データキューブ (数値をセルの値として持つデータキューブなので数値データキューブと呼ぶ) によって、「ページ別アクセス数 対 時間」や「ユーザ種別アクセス数 対 時間」など、自由に決めた属性の組についての数値グラフを即、計算する。そして、このグラフ表示を元に、着目する状況を選び、そこでどのような事象の組み合わせが生じているかを、このデータマイニングサーバを使って即座にアイテムセット分析により計算する。図 1 の例でいうと、ある数値グラフにおいて変化の起きている箇所で生じる現象の内容を得たり (図 1 左下)、特定の種類のユーザの行動を時間軸に沿って追跡する（図 1 右下)、などが可能になる。

このシステムでは、当然、従来の OLAP と同じく、多様な属性の組や概念階層に応じた高頻度アイテムセット計算を高速に行う必要がある。 この要求に答えるため、著者らは、時間やイベント種別といった条件によって場合分けされた多次元空間を用意し、その各セルに当該レコード集合から計算した高頻度アイテムセットを格納したデータキューブモデル「アイテムセットキューブ」を提案する。そして、これを用いて、スライスやロールアップに相当する変形演算を施し、問い合わせの指定した状況における高頻度アイテムセットを計算する。

本モデルの定義や演算処理の技法は文献 [2], [3] に述べてきた。本稿では、2 節でこれらの概要を述べた後、3 節で実体化とロールアップ処理の算法を示し、4 節で実データ (Web アクセスログ) の適用例を述べていく。

2. アイテムセットキューブによる多次元ログ分析の概要

2.1 前　　提

以下、Web サーバのアクセスログ分析を例にとってアイテムセットキューブを概説する。

アクセスログ分析では、ログ列は「同一ユーザから一定時間内に連続してアクセスされたページの列」を表すセッションレコードに変形される。以下、本稿では、セッションレコードを、[ユーザドメイン、アクセス開始時刻、アクセスの発生した月、アクセスされたページ集合] の 4 属性で表現する。そして、ユーザドメイン (の種類)、アクセスの生じた月 (1 月、2 月、…)、アクセスされたページ集合、の 3 属性に着目したログ分析を考える。以下、レコードという時は、セッションレコードを指す。

次に、各属性を 1 次元として考えて多次元空間を作り、各次元をいくつかの区切りに分割してデータキューブ構造を作ってレコードを分類する。

例として、「ユーザドメイン」属性を、ドメインの種類 (ac.jp,co.jp, .com 等) によって $\{G_1, G_2, G_3\}$ の 3 つに分割する。1 レコードは相異なる G_i, G_j に同時に属すことはない。このようにレコードを排他的に分ける分割 (この例では $\{G_i\}(i = 1, 2, 3)$) のことを、**排他的な分割**と呼ぶ。各値 G_i を、分割における**区切り**と呼ぶ。「月」属性は、1 月から 12 月に分割する。

一方、Web ページは、学会などのコミュニティサイトであれば、大会 (E_1) や講演会 (E_2)、会員申込み (E_3) などの行事に関するページに分類される。 ページ p が E_i に分類される時、E_i を p のページ種類と呼ぶ。 これに従って言うと、「(1 レコードでアクセスされた) ページ集合」属性は、述語「ページ種類 E_i に属すページが当該レコードでアクセスされている」(i=1,2,3) で分割される。(この述語自体も以下では記号 E_i で参照する)。ここで、1 レコードは複数のページをアクセスするから、結局、1 レコードは相異なる E_i, E_j に属す可能性がある。このような分割を**非排他的な分割** と呼ぶ。この時、概念的には、当該レコードは照応する各 E_i へコピーされて分類されると考える。

2.2 従来手法:数値データキューブ

ログ分析を多次元的に行う手法の主流は、データキューブを

使った OLAP である [14] [9]。図 2 左上に、「月」,「ユーザドメイン」、「ページ集合」の 3 次元からなる数値データキューブを示す. キューブの各次元 X は、それに与えられた分割 $P_x = \{x_1, x_2, \ldots, x_{k_x}\}$ により区切られる。キューブの最小構成要素であるセルは、各次元 X から区切り 1 つ (これを x_i で表す) を選んだ時に、これら x_i の論理積を全次元について行った論理式として定義される。各セルは、その論理式を満たすレコードの総数 (ヒット数) を格納する。　OLAP では、このようなデータキューブをあらかじめ計算しておき、利用者の問い合わせに対して、該当するセルの値を集計して返す。問い合わせは、「4 月におけるページ種類 E_1、E_2、E_3 別のヒット数を求めよ」というように、必ずしもデータキューブの構造とは一致しない。そのため、キューブから必要な部分 (4 月に関するセル全て) をスライスで取り出し、必要でない属性（この問い合わせでは、「ユーザドメイン」) を「全ユーザ」へとロールアップして結果を返す。

図 2 は、より簡単な問い合わせ「ページ種類別のヒット数を月単位で求めよ」を処理する場合を示す。この場合、図中、不要な次元である「ユーザドメイン」を全ユーザ (図中の G:All) へロールアップし、元の数値データキューブを変形する。この処理自体は単純な集計処理である。得られた結果をグラフ表示すると図 2 右下になる。

このとき、講演会ページへのヒット数が 6 月に急増していることに着目すると、そこでのユーザ種類別のアクセス数内訳を求めることが考えられる。これがドリルダウン処理であり、元のキューブからスライス、ロールアップにより計算する. 図 2 左下が得られる結果である。

2.3　数値データキューブとアイテムセット分析の組み合わせの問題点

上記のように、従来の数値用データーキューブでは、問い合わせに応じて、スライスとロールアップにより、事前に実体化したキューブを変形する。ロールアップ時には元のログデータの再スキャンを行わないため、多様な問い合わせに対しても高速に応答できる。

一方で、データキューブにより任意の多次元空間上の数値グラフを即計算できる結果として、そのグラフにおける任意の指定状況においてどのような事象の組み合わせが多いのか、をも同時に調べたいという要求が考えられる。図 2 であれば、

Q1: 「申込ページを見たレコード集合において、良く見られるページの組み合わせは何か。それは 4、5、6 月と時間が変わるとどう変化するか。」

や、

Q2: 「講演会ページを見たレコード数は 6 月に急増している。そのユーザ別アクセス数を計算した時、各ユーザグループ G_1、G_2、G_3 が見た頻出ページ組は何か」

という問い合わせである。

図 3 は、これらのデータマイニングを求める問い合わせを、OLAP と併せて発行した状況を示している。図中、右下部が Q1、左下が Q2 に対応する。

上記のような分析を行うには、着目した状況を満たすレコード集合をまず求め、そこから、調べたい対象となる事象を 1 アイテムとして、高頻度アイテムセットを計算するしかない。Web ログデータ分析の場合、調べたい事象は、「どのページをアクセスしたか」であるので、ページを 1 アイテムとして、同一レコードに高頻度に出現するアイテムセットを求めることになる。Apriori [11] や FP-growth などの列挙アルゴリズムがあるが、計算 1 回あたりのメモリ消費量や CPU 時間の高い技法である。

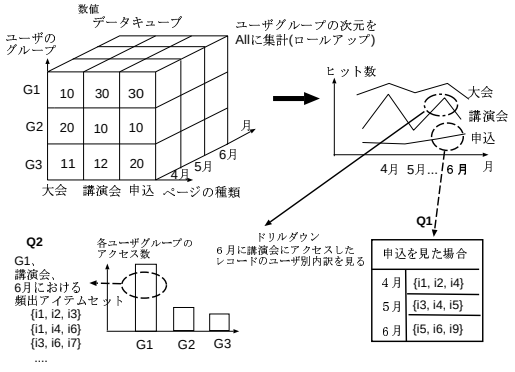


図 3　数値データキューブと併せたアイテムセット計算機構

一方、図 3 においてアイテムセット計算を引き起こす前提となった数値グラフの計算は、OLAP では、他の属性の組み合わせ、例えば、「ユーザ種別アクセス vs. 時間軸」や、「1 年間累計でのユーザ種別アクセス vs. ページ種類別アクセス」についても発生しうる。また、各次元の分割の概念階層の指定も多様である。例えば、時間軸を月単位で計算するか、四半期単位で計算するかである。従って、図 3 のようにデータマイニングを OLAP と組み合わせても、着目状況ごとにレコード集合を導出してアイテムセット計算を毎回行う実装手法では、OLAP と同程度の自由な多次元データ分析に対応した効率的なデータマイニングサーバは実現できない。

2.4　提案：アイテムセットキューブ

前節で述べた問題に対して、著者らは、高頻度アイテムセットを各セルの値として持つようなデータキューブモデルを提案している。これを「アイテムセットキューブ」と呼ぶ。アイテムセットキューブは、2.1 節の前提と用語の下で、以下のように定義される。

1. データキューブの各セルは、従来の数値データではなく、そのセルの条件を満たすようなレコード集合から計算された高頻度アイテムセットを持つ。ここでいう高頻度アイテムセットとは、支持率 s の下であるアイテムセット I がセル c において、I の出現回数が c における条件を満たすレコード総数 N_c に対して、そのセルで $N_c \times s$ 回以上である場合を指す。s はキューブを構成する全てのセルにおいて共通の定数として与える。[]

アイテムセットキューブをログデータから計算する演算を実体化と呼ぶ。ロールアップとドリルダウンも、従来の数値キューブと同様の変形操作として定義される。(形式的定義は文献 [2] を参照)。 つけくわえておくと、アイテムセットキューブのロールアップなどの変形操作は、セルの値が支持率 s での高頻度ア

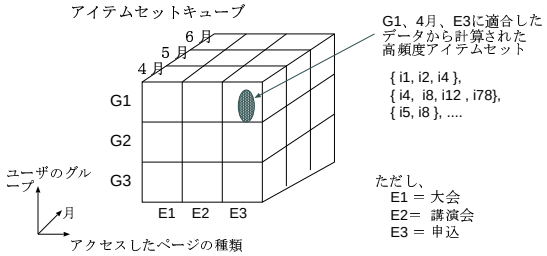


図 4　アイテムセットキューブ

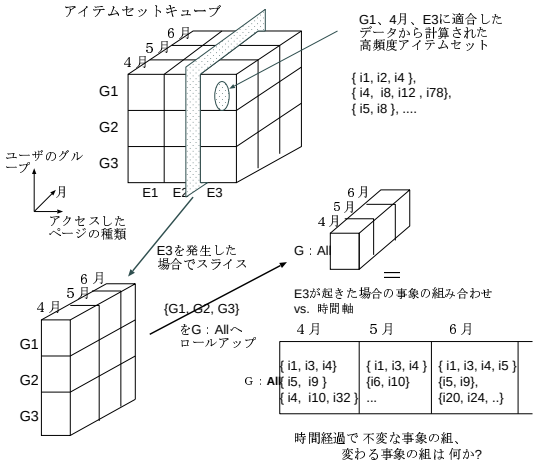


図 5　アイテムセットキューブの変形例

イテムセットであるため、数値データキューブのような単純な集計処理では実現できない。

図 4 は、ユーザドメイン (G_1, G_2, G_3 の 3 グループ)、月 (1-12 月)、アクセスされたページの種類 (「E_i に関するページがアクセスされた」、E_1=大会、E_2=講演会、E_3=申込) の 3 属性で実体化したアイテムセットキューブの例である。

図 5 は、図 4 のキューブを用いて、2.3 節で述べた問い合わせ Q1 =「E_3 に関するページがアクセスされた場合において生じる高頻度アイテムセットを、月別に示せ」を処理する場合である。すなわち、「ページの種類= E_3」でスライスを行い、「ユーザドメイン」次元を「全ユーザ」(G:All) へロールアップすることで問い合わせに応じた結果を計算する。

このように、あらかじめ必要な分割を設定しアイテムセットキューブを実体化しておき、問い合わせに応じてスライス、ロールアップすることで、オンライン的なアイテムセット分析ができる。

例： 先の問い合わせ Q1 を一般化し、「各ページ種類 E_i について、それに属すページをアクセスしたユーザが良く見たページ集合を、月別に出せ。(ただし、E_i=大会、講演会、申込)」という問い合わせを考えよう。この処理は、数値データキューブの場合に行ったスライス, ロールアップの演算系列と同じ操作列 (この場合は「ユーザグループを G:All にロールアップ」) を図 4 のアイテムセットキューブに適用することで、当該する高頻度アイテムセット集合を得ることができる。問い合わせ Q2 も、同様に、元のアイテムセットキューブから適切なスライスとロールアップにより計算される。

2.5　技術的課題

アイテムセットキューブの技術的課題は次の通り:

*1. 実体化が効率的にできること。

*2. 概念階層の有無によらず、ロールアップが効率的に実現可能であること。

以下、この 2 点の技法を説明する。

3.　アイテムセットキューブの演算

3.1　実　体　化

アイテムセットキューブの実体化処理は，属性の分割が排他的な分割である場合と，非排他的な分割である場合に分けられる．属性の分割が排他的な分割である場合，セルに対応するレコード集合は重複しないため，セルごとに個別に Apriori を用いて実体化をすればよい．この実体化の方法を Naive 法と呼ぶ.

しかし，属性の分割が非排他的な分割である場合，セルに対応するレコード集合は重複するため，Naive 法で実体化すると，セルごとの候補アイテムセットが重複する場合があるため，同じ候補アイテムセットに対して複数回データベーススキャンを行うことになる．そこで，1 度のデータベーススキャンで，さらに重複する候補アイテムセットを除去して実体化することで，実体化の高速化をはかる．この方法を Cubic-Apriori(CA 法) と呼ぶ.

- **実体化アルゴリズム Cubic-Apriori**

1 次元の非排他的な分割により規定されたセル $C_1, C_2, ..., C_n$ を支持率 s で実体化する.

今，全てのアイテムセット I に，各セル $C_i (i = 1, ..., n)$ での出現回数を c_i として，$v(I) = [c_1, c_2, ..., c_n]$ を与える.

step1.　各レコード r について長さ 1 のアイテムセット I_1 の数え上げを行い，r が満たすセル C_i についてのみ c_i をインクリメントする．全レコードスキャン後，セルあたりの頻度足切りを行い，長さ 1 の高頻度アイテムセットの集合 L_1 を確定する.

step2.　長さ ($\geqq 2$) の候補アイテムセット I_k を長さ $k-1$ の高頻度アイテムセット L_{k-1} から生成する．I_k が候補となるのは，あるセルにおいて，I_k を構成する長さ $k-1$ のアイテムセット全てが高頻度の時に限られる.

step3.　各レコード r について，候補アイテムセット I_k について r が満たすセル C_i についてのみカウンタ c_i をインクリメントする．全レコードスキャン後に，L_k を確定する.

step4.　L_k が空になるまで，step2 に戻る.

3.1.1　実体化の定量評価

実体化アルゴリズムに用いられている CA 法は，セルごとに個別にアイテムセット計算しない．そのため，非排他的な分割をもつ属性を実体化を行う際に効果を発揮する．ここでは，実体化アルゴリズムの評価を行う．計算機環境・人工データベースの概要は次のとおりである.

・計算機環境

CPU：Pentium4 2.40GHz

RAM：1024MB

OS：Vine Linux 2.6CR

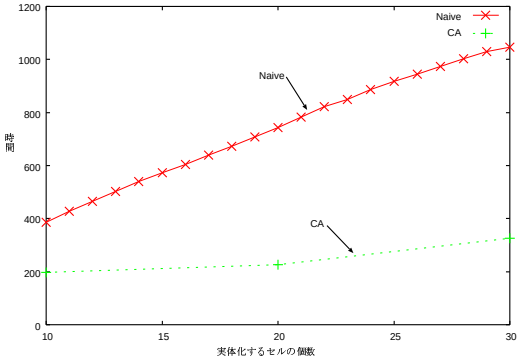


図 6　実体化計算における 1 次元のセル数と計算時間

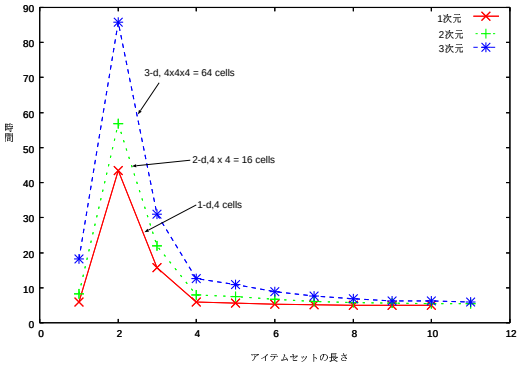


図 7　1，2，3 次元におけるアイテムセット長の計算時間比較

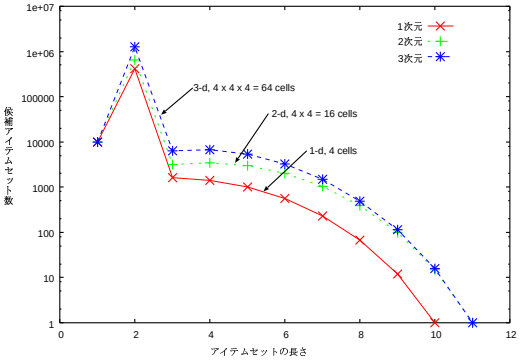


図 8　1，2，3 次元におけるアイテムセット長ごとの候補アイテムセット数

・人工データベース

最大アイテム番号 (アイテム数)：10000
総レコード数：300000
1 レコード内のアイテム数の平均：20
高頻度アイテムセットの長さの平均：4

また，実体化するセルの内容は，レコードの重複を多くするため，第 i 番目のセルの定義 C_i ＝「アイテム番号 $300i$〜$300i + 299$ のうちどれかのアイテムを含む」とし，30 個のセルを定めた．なお，実体化を行うにあたり，全セルの閾値率は，1%とした．図 6 は，非排他的な分割でセルが設定された場合における，1 次元のアイテムセットキューブを，Naive 法と CA 法で実体化した計算時間である。図 6 から，CA 法では実体化するセルの数が増加したとしても，計算時間はあまり増加していない．このことよりセル数の増加に対して CA 法が冗長計算を回避できることを示している．

2 次元，3 次元の場合:非排他的な分割をもつ属性からなる 2 次元，3 次元の場合でも，実体化を行うときは該当する多次元上のセルを列挙し 1 次元のセル列と考えることで，CA 法をそのまま適用できる．

図 7 は，上記で使用したセルを 1 次元の時に 4 セル（C_0〜C_3），2 次元の時に 16 セル（C_0〜$C_3 \times C_{10}$〜C_{13}），3 次元の時に 64 セル（C_0〜$C_3 \times C_{10}$〜$C_{13} \times C_{20}$〜C_{23}）といったように，1 次元増えるごとに 4 セルをかけていってセル数を増やしていった場合の実体化の計算時間である．

また，図 7 のように，次元を増やした場合であっても，セルを 1 次元のセル列として CA を適用することで，高速な実体化が可能である．

3.2　ロールアップ

ロールアップを行う場合，排他的な分割を与えた属性をロールアップする場合と，非排他的な分割を与えた属性をロールアップする場合がある．前者においては，相異なるセルに属するレコード集合は共通集合をもたない．このような場合は，各セルの高頻度アイテムセットの集合をマージして作ったアイテムセット集合のみが，ロールアップ後の候補アイテムセットとなる． なぜなら，共通レコードをもたないレコード集合 $D_1, D_2, ..., D_n$ の和 $D_1 \cup D_2 \cup ... \cup D_n$ において，高頻度であるアイテムセット I は，必ず $D_i (i = 1, 2, ..., n)$ のいずれかで高頻度アイテムセットとなる．

逆に，非排他的な分割を与えた属性については，上記は成立しない．この場合のロールアップについては, [2] にて述べられているとおり，分割に概念階層が与えられている場合に，下位階層のデータキューブを計算する時と同時に上位階層のキューブを計算し，CA 法の特性により計算負荷を削減することで，高速化を行っている．

排他的な分割を与えた属性をロールアップする場合は，ロールアップ対象となるセルに含まれる高頻度アイテムセットをマージする必要がある．そして，全てのセルの高頻度アイテムセットをマージした後，それらを候補アイテムセットとして，1 回データベーススキャンを行い，ロールアップ後のセルに適した高頻度アイテムセットを決定するだけで良い．

3.3　ロールアップの定量評価

ロールアップの性能を調べるため，人工データを用いた定量評価を行った．実験で用いた人工データの詳細は次のとおりである．

・人工データ 1〜10 の 1 つあたりの詳細

アイテム数：10000
総レコード数：30000
1 レコード内のアイテム数の平均：20
高頻度アイテムセットの長さの平均：4

実験では，人工データ 1〜10 を非排他的な分割をもつセル 5 個 (実体化の実験で用いたセル C_0〜C_4) を条件として実体化を行った結果のファイルを 10 個をロールアップする場合と，人工データ 1〜10 をマージし，30 万件のデータを実体化する場合

の処理時間を比較した．なお，ロールアップを行う場合も，実体化を行う場合も，全セルの閾値率は 1%とした．

図 9 は実行結果である．図 9 からわかるとおり，条件を変更して再実体化を行う場合の処理時間と比べ，ロールアップ処理の時間は 4 分の 1 程度である．また，図 9 から，候補アイテムセット数の最大値が再実体化を行う場合と比べて，ロールアップ時は 200 分の 1 であることがわかる．この特性により，メモリの使用量を削減することができるため，オンライン上のマルチユーザの問い合わせにも，利用することができる．

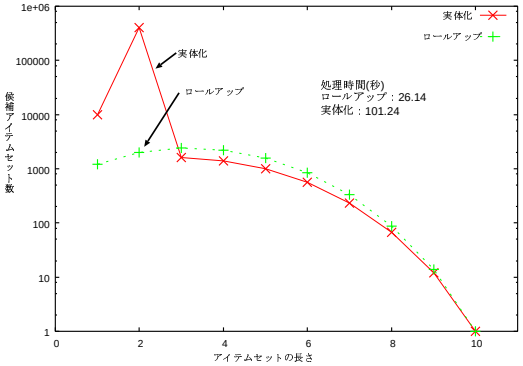


図 9　候補アイテムセット数

4.　実データへの適用

この節では，実データを用いてアイテムセットキューブによる分析の有効性を示す．使用するデータは，ある会員制団体の Web サイトのアクセスログ 2002 年分で，ログ数は 213863 行である．このデータから，検索エンジン等の巡回ロボットのアクセスログを除去し，セッションレコードに変形をして分析を行う．セッションレコードの件数は，17293 件であった．

4.1　分　　析

数値のキューブ，アイテムセットキューブの実体化は，共に次の 3 次元で行った：ドメイン種別，月，イベント種類．
なお，各次元に与えられた属性の分割は次のとおりである．(図 10)

- ドメイン種別：ac ドメイン，co ドメイン，com ドメイン，その他のドメイン
- 月：1 月～12 月
- イベント種類：講演会のいずれかを見た，大会のいずれかを見た，申込みを見た，all(前出のいずれかに属する)

図 11 は，数値データキューブのユーザドメインを全ユーザへロールアップで集計し，「イベント種類」と「月」の 2 次元にして表示した数値グラフである．図 11 からは，大会を見たレコード群の 7 月，10 月，11 月，12 月分のヒット数が他の月と比べて，増加していることがわかる．逆に講演会を見たレコード群は 7 月はあまり増加せずに，10 月～12 月は急増していることがわかる．また，申込みを見たレコード群は他 2 つのイベント種類と比べて，緩やかな増減を示している．ここでヒット数の増減の大きい，大会を見たレコード群と講演会を見たレコード

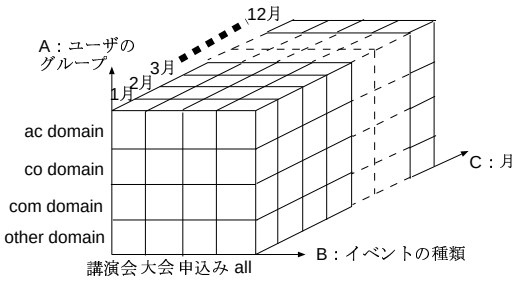


図 10　上記の条件により実体化されたキューブ

群の動作を調べるため，時間軸を月別から四半期 (3ヶ月ごと) にロールアップとユーザグループを全ユーザのロールアップ 2 つの演算を図 10 アイテムセットキューブに適用して，四半期別、イベント種類ごとの高頻度アイテムセットをサポート数順上位 20 個出力した (図 12)．図 12 を見ると，前半の第 1，第 2 四半期 (1Q，2Q) では，開催されたイベントが少ないこともあり，大会を見た人と講演会を見た人のレコード群は、それぞれ異なる動作をしている．対して，開催されたイベントが多い後半の第 3，第 4 四半期 (3Q，4Q) では，どちらのレコード群も類似した動作を示していることがわかる (図 13，図 14)．

次に，図 11 の楕円で囲った区域で，講演会を見たレコード群が，7 月，9 月から 12 月に大きく増加し，それに合わせて all も増加していることに注目し，この時期に講演会を見たレコード群がどのようなページを組み合わせて見ているかを調べる．図 14 からは，講演会を見たレコード群が 4Q では kouenkai23, taikai24 と組み合わせて，datamining や oracle02oct3 といったこの時期の特別なイベントを見ていることがわかる．

さらに，講演会を見たレコード群の動作の詳細を見るためにユーザドメイン種別の次元を追加する．数値のデータキューブを講演会を見たレコード群，3Q，4Q の条件でドリルダウンした結果 (図 15 上部) から，com ドメインと other ドメインが講演会を見たグループの大半を占めていることがわかる．そこで，other ドメインと com ドメインの動作を調べるため，アイテムセットキューブをドリルダウンした結果が図 15 下部である．図 15 下部から，other ドメインは図 14 と似た動作をとっていることがわかった．対して，com ドメインは，ほとんど違う動作をとっていることがわかる．これらのことより，9 月～12 月の講演会を見たレコード群の増加は，other ドメインが主要因であることがわかった．

図 16 は，分析する次元を変更した例である．図 16 では，数値のデータキューブのイベント種類の軸をロールアップで all に集計し，「ユーザ種別」と「月」の 2 次元で表示したものである．このように，分析条件を変更した場合でも，アイテムセットキューブに同じ演算列を適用することで，各ユーザが行った主要動作を時間順に得ることができる (図 17)．図 17 からは，年間を通して，ac と co と other ドメインは類似した動作をしているのに対して，com ドメインは異なる動作をとっていることがわかる．このことから，other ドメインには人間的な動作をするユーザが多く含まれているが，com ドメインには，Web 巡回ロボットのような機械的な動きをするユーザが多く含まれて

いることがわかる．

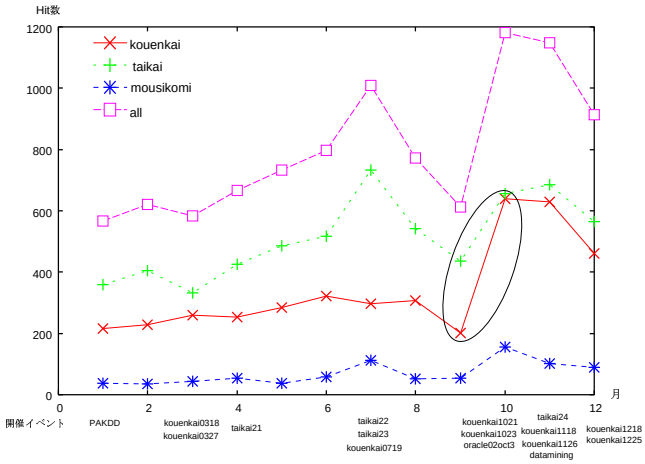


図 11　各イベント種類の月別ヒット数

Top20 Itemset

	1Q	2Q	3Q	4Q
kouenkai	hit 704 Events, kouenkai20020318, index sup 3% Events, index, koukai sup 3% kouenkai20020318, index, koukai sup 2% VLDB2001, kouenkai20020318, index sup 2% Events, kouenkai20011207, index sup 2% VLDB2001, index, koukai sup 2% Events, kouenkai20020318, koukai sup 2% Events, Houjin, index sup 2% VLDB2001, Events, koukai sup 2% Events, index, sibukiyaku sup 2%	hit 859 Events, taikai21, index sup 2% Events, index, koukai sup 2% kouenkai20020327, taikai21, index sup 2% Events, kouenkai20020327, index sup 2% Events, index, sibukiyaku sup 2%	hit 804 Events, index, koukai sup 4% Events, kouenkai20020719, koukai sup 4% Events, taikai23, koukai sup 3% Events, kouenkai20020719, taikai23 sup 3% Events, kouenkai20020719, index sup 3% Events, kouenkai20020719, links sup 3% Events, taikai22, koukai sup 3% Events, taikai23, index sup 3% Events, taikai23, links sup 3% Events, koukai, links sup 3% taikai22, taikai23, koukai sup 3% taikai22, koukai, mailing sup 3% Events, taikai22, taikai23 sup 3% Events, taikai23, sibukiyaku sup 3% Events, koukai, sibukiyaku sup 3% kouenkai20020719, index, koukai sup 3% taikai22, taikai23, links sup 3% Events, kouenkai20020719, taikai22 sup 3% taikai22, taikai23, links sup 3% Events, taikai23, mailing sup 3%	hit 1729 datamining, kouenkai20021023, index sup 4% Events, kouenkai20021023, index sup 3% kouenkai20021023, taikai24, index sup 3% Events, index, sibukiyaku sup 2% Houjin, index, sibukiyaku sup 2% Events, Houjin, sibukiyaku sup 2% kouenkai20021023, index, register sup 2% Events, index, koukai sup 2% index, register, sibukiyaku sup 2% Events, oracle02oct3, index sup 2% datamining, Events, index sup 2% datamining, kouenkai20021023, taikai24 sup 2% Events, kouenkai20021126, index sup 2% datamining, Events, kouenkai20021023 sup 2% kouenkai20001116, taikai18, VLDB2000 sup 2% Events, oracle02oct3, koukai sup 2% Events, index, links sup 2% Events, koukai, sibukiyaku sup 2% Events, kouenkai20021023, taikai24 sup 2%
taikai	hit 1097 Events, index, koukai sup 2%	hit 1429 Events, taikai21, index sup 6% Events, index, koukai sup 3% taikai21, index, koukai sup 3% taikai21, index, sibukiyaku sup 3% VLDB2001, taikai21, index sup 2% Events, taikai21, koukai sup 2% VLDB2001, Events, index sup 2% Houjin, index, sibukiyaku sup 2% Dlib, taikai21, index sup 2% VLDB2001, Events, taikai21 sup 2% Events, taikai21, enquete sup 2% taikai21, enquete, index sup 2% Events, enquete, index sup 2% VLDB2001, index, koukai sup 2% Events, index, sibukiyaku sup 2% VLDB2001, Events, koukai sup 2% VLDB2001, taikai21, koukai sup 2% Events, taikai21, sibukiyaku sup 2% taikai21, index, register sup 2% index, register, sibukiyaku sup 2%	hit 1710 Events, index, koukai sup 3% taikai23, index, koukai sup 3% Events, taikai23, index sup 3% taikai22, taikai23, index sup 3% Events, taikai23, koukai sup 2% taikai23, index, links sup 2% Dlib, taikai23, index sup 2% taikai22, taikai23, koukai sup 2% taikai22, taikai23, links sup 2% taikai22, index, koukai sup 2% taikai23, index, sibukiyaku sup 2% index, register, sibukiyaku sup 2% taikai23, Houjin, index sup 2% taikai23, index, register sup 2% Events, taikai23, links sup 2% Events, taikai22, koukai sup 2% taikai22, taikai23, sibukiyaku sup 2% taikai23, koukai, sibukiyaku sup 2% Dlib, Events, index sup 2% Dlib, taikai23, koukai sup 2%	hit 1907 index, register, sibukiyaku sup 3% Houjin, index, sibukiyaku sup 2% datamining, taikai24, index sup 2% Events, taikai24, index sup 2% Events, Houjin, index sup 2% kouenkai20021023, taikai24, index sup 2% Events, index, koukai sup 2% Events, index, sibukiyaku sup 2% Dlib, Events, index sup 2% datamining, Events, index sup 2% Events, Houjin, sibukiyaku sup 2% index, mailing, sibukiyaku sup 2% Dlib, index, sibukiyaku sup 2%

図 12　各イベントの四半期別ヒット数

2002年 閾値2% 2item

	1Q	2Q	3Q	4Q
大会	hit 1097 ⋮	hit 1429	hit 1710	hit 1907

	3Q	4Q
大会	hit 1710 Events, index, koukai sup 3% taikai23, index, koukai sup 3% Events, taikai23, index sup 3% taikai22, taikai23, index sup 3% Events, taikai23, koukai sup 2% taikai23, index, links sup 2% Dlib, taikai23, index sup 2% taikai22, taikai23, koukai sup 2% taikai22, taikai23, links sup 2% taikai22, index, koukai sup 2% taikai23, index, sibukiyaku sup 2% index, register, sibukiyaku sup 2% taikai23, Houjin, index sup 2% taikai23, index, register sup 2% Events, taikai23, links sup 2% Events, taikai22, koukai sup 2% taikai22, taikai23, sibukiyaku sup 2% taikai23, koukai, sibukiyaku sup 2% Dlib, Events, index sup 2% Dlib, taikai23, koukai sup 2%	hit 1907 index, register, sibukiyaku sup 3% Houjin, index, sibukiyaku sup 2% datamining, taikai24, index sup 2% Events, taikai24, index sup 2% Events, Houjin, index sup 2% kouenkai20021023, taikai24, index sup 2% Events, index, koukai sup 2% Events, index, sibukiyaku sup 2% Dlib, Events, index sup 2% datamining, Events, index sup 2% Events, Houjin, sibukiyaku sup 2% index, mailing, sibukiyaku sup 2% Dlib, index, sibukiyaku sup 2%

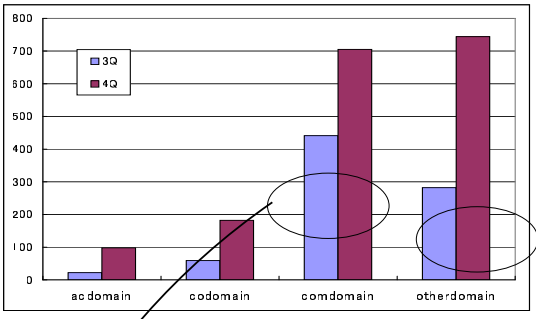
図 13　大会をみたレコード群の第 3，第 4 四半期の高頻度アイテムセット (2itemset)

2002年 閾値2% 2item

	1Q	2Q	3Q	4Q
講演会	hit 704 ⋮	hit 859	hit 804	hit 1729

	3Q	4Q
講演会	hit 804 kouenkai20001116, taikai18 sup 7% Events, index sup 6% Events, koukai sup 5% index, koukai sup 5% Events, kouenkai20020719 sup 5% kouenkai20020719, index sup 5% Events, taikai23 sup 4% kouenkai20020719, koukai sup 4% Events, sibukiyaku sup 4% taikai23, index sup 4% Events, links sup 4% taikai22, koukai sup 4% taikai23, koukai sup 4% kouenkai20020719, links sup 4% Events, Houjin sup 3% taikai22, taikai23 sup 3% kouenkai20020719, taikai23 sup 3% koukai, mailing sup 3% mailing, sibukiyaku sup 3% Dlib, Events sup 3% Events, taikai22 sup 3% kouenkai20020719, taikai22 sup 3% kouenkai20020719, mailing sup 3% taikai22, mailing sup 3% koukai, links sup 3%	hit 1729 datamining, index sup 10% Events, index sup 10% kouenkai20021023, index sup 10% oracle02oct3, index sup 7% index, sibukiyaku sup 5% index, register sup 5% Houjin, index sup 5% Events, oracle02oct3 sup 5% index, koukai sup 5% datamining, Events sup 5% taikai24, index sup 5% index, links sup 5% Events, kouenkai20021023 sup 4% Events, sibukiyaku sup 4% Events, Houjin sup 4% Events, koukai sup 4% datamining, kouenkai20021023 sup 4% datamining, taikai24 sup 4% kouenkai20001116, taikai18 sup 4% oracle02oct3, koukai sup 3% Houjin, sibukiyaku sup 3% register, sibukiyaku sup 3% oracle02oct3, Houjin sup 3% kouenkai20021023, taikai24 sup 3% Events, links sup 3%

図 14　講演会を見たレコード群の第 3，第 4 四半期の高頻度アイテムセット (2itemset)



	3Q	4Q
other ドメイン	hit 282 Events, kouenkai20020719, koukai sup 8% Events, index, koukai sup 7% Events, kouenkai20020719, index sup 6% Events, taikai23, koukai sup 6% Events, koukai, links sup 6% Events, mailing, sibukiyaku sup 6% Events, kouenkai20020719, taikai23 sup 6% Events, kouenkai20020719, links sup 6% Events, Houjin, sibukiyaku sup 6% Events, koukai, sibukiyaku sup 6% Events, sibukiyaku, links sup 6% taikai22, koukai, mailing sup 6% Events, kouenkai20020719, sibukiyaku sup 5% Events, taikai22, koukai sup 5% Events, taikai23, mailing sup 5% Events, taikai23, sibukiyaku sup 5% kouenkai20020719, taikai23, koukai sup 5% kouenkai20020719, index, koukai sup 5% taikai22, taikai23, koukai sup 5% taikai22, taikai23, mailing sup 5% taikai23, koukai, mailing sup 5% taikai23, mailing, sibukiyaku sup 5%	hit 744 datamining, Events, index sup 6% Events, index, sibukiyaku sup 5% Events, Houjin, sibukiyaku sup 5% Events, Houjin, index sup 5% Events, oracle02oct3, index sup 5% Houjin, index, sibukiyaku sup 5% Events, kouenkai20021023, index sup 5% index, register, sibukiyaku sup 5% datamining, kouenkai20021023, index sup 4% Events, oracle02oct3, koukai sup 4% Events, oracle02oct3, links sup 4% Events, index, koukai sup 4% datamining, taikai24, index sup 4% Events, oracle02oct3, Houjin sup 4% Events, oracle02oct3, sibukiyaku sup 4% Events, sibukiyaku, links sup 4% kouenkai20001116, taikai18, VLDB2000 sup 4% Events, mailing, sibukiyaku sup 4% oracle02oct3, index, koukai sup 4% Dlib, Events, Houjin sup 4% Dlib, Events, sibukiyaku sup 4% kouenkai20021023, taikai24, index sup 4%
com ドメイン	hit 405 Events, koukai sup 2% index, koukai sup 2% register, sibukiyaku sup 2%	hit 705 Events, taikai23 sup 2% Events, Whoswho sup 2% oracle02oct3, Whoswho sup 2% Dlib, oracle02oct3 sup 2%

図 15　講演会を見たレコード群で，かつ第 3，第 4 四半期におけるドメイン別ヒット数とそのアイテムセット

5. 終わりに

本稿では，膨大なログを理解するために多次元的なデータマ

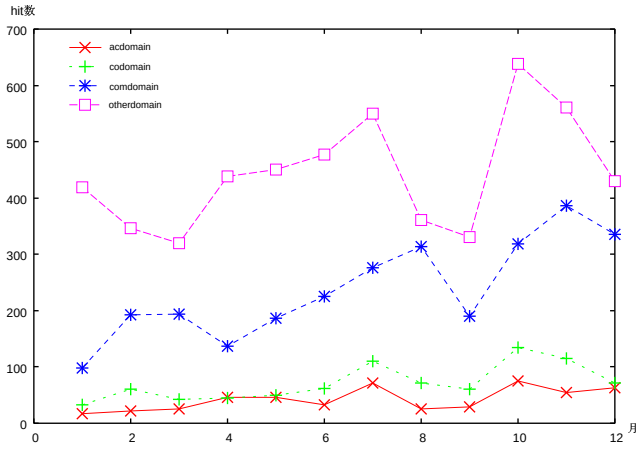


図 16　2002 年ドメイン別ヒット数

2002年 閾値2%

	1Q	2Q	3Q	4Q
other ドメイン 2item set	hit 1085 Events, index sup 5% index, register sup 4% index, sibukiyaku sup 4% index, koukai sup 4% register, sibukiyaku sup 3% kouenkai20020318, index sup 3% Events, koukai sup 3% VLDB2001, index sup 2% Houjin, index sup 2% Dlib, index sup 2% Events, taikai20 sup 2% VLDB2001, koukai sup 2% taikai20, index sup 2% Houjin, register sup 2% Houjin, sibukiyaku sup 2% koukai, sibukiyaku sup 2% index, mailing sup 2% koukai, register sup 2% mailing, sibukiyaku sup 2% Dlib, register sup 2%	hit 1366 taikai21, index sup 9% Events, index sup 7% Events, taikai21 sup 6% index, sibukiyaku sup 6% index, register sup 5% index, koukai sup 4% VLDB2001, index sup 3% Events, koukai sup 3% Houjin, index sup 3% Dlib, index sup 3% register, sibukiyaku sup 3% taikai21, sibukiyaku sup 3% taikai21, koukai sup 3% Houjin, sibukiyaku sup 3% VLDB2001, Events sup 3% enquete, index sup 2% VLDB2001, taikai21 sup 2% taikai17, index sup 2% Dlib, taikai21 sup 2% Events, enquete sup 2%	hit 1242 taikai23, index sup 10% index, register sup 7% Events, index sup 6% index, sibukiyaku sup 5% register, sibukiyaku sup 4% taikai22, index sup 4% index, koukai sup 4% index, links sup 4% kouenkai20001116, taikai18 sup 4% Events, koukai sup 4% taikai22, taikai23 sup 3% Houjin, index sup 3% Events, taikai23 sup 3% taikai23, koukai sup 3% Dlib, index sup 3% taikai23, sibukiyaku sup 3% taikai23, links sup 3% Houjin, sibukiyaku sup 3% VLDB2001, index sup 3% Events, sibukiyaku sup 3%	hit 1629 Events, index sup 9% index, register sup 9% index, sibukiyaku sup 8% datamining, index sup 7% kouenkai20021023, index sup 7% register, sibukiyaku sup 6% Houjin, index sup 6% taikai24, index sup 5% index, koukai sup 4% index, links sup 4% oracle02oct3, index sup 4% Houjin, sibukiyaku sup 4% Events, sibukiyaku sup 4% index, mailing sup 4% Dlib, index sup 4% Events, Houjin sup 4% datamining, Events sup 3% Events, kouenkai20021023 sup 3% Events, koukai sup 3% kouenkai20001116, taikai18 sup 3%
co ドメイン 2item set	hit 136 Events, taikai20 sup 8% index, sibukiyaku sup 8% Events, index sup 8% kouenkai20020318, index sup 8% index, koukai sup 7% index, register sup 6% VLDB2001, index sup 5% Events, kouenkai20020318 sup 5% taikai16, index sup 5% Houjin, index sup 5% Dlib, index sup 4% taikai19, index sup 4% VLDB2000, index sup 4% register, sibukiyaku sup 4% VLDB2001, koukai sup 3% Events, Whoswho sup 3% Events, koukai sup 3% kouenkai20010424, index sup 3% taikai17, index sup 3% Whoswho, index sup 3%	hit 157 Events, index sup 6% taikai21, index sup 6% index, sibukiyaku sup 5% taikai16, index sup 4% taikai17, index sup 4% taikai19, index sup 4% Events, taikai21 sup 3% Houjin, index sup 3% index, register sup 3% taikai16, taikai19 sup 3% taikai19, taikai21 sup 3% Houjin, sibukiyaku sup 3% VLDB2000, index sup 3% register, sibukiyaku sup 3% Dlib, Events sup 2% Dlib, taikai21 sup 2% Dlib, index sup 2% SIGMOD, index sup 2% Events, taikai17 sup 2% Events, shortmemo1 sup 2%	hit 242 taikai23, index sup 13% index, register sup 11% Houjin, index sup 9% index, sibukiyaku sup 9% index, koukai sup 8% Dlib, index sup 7% Events, index sup 7% register, sibukiyaku sup 7% index, links sup 7% taikai23, Houjin sup 6% taikai22, index sup 6% index, mailing sup 6% Houjin, register sup 5% Dlib, taikai23 sup 5% taikai16, index sup 5% Houjin, sibukiyaku sup 5% taikai13, index sup 4% taikai20, index sup 4% Houjin, mailing sup 4% taikai21, index sup 4%	hit 357 Events, index sup 10% index, register sup 8% taikai24, index sup 7% Houjin, index sup 7% datamining, index sup 7% kouenkai20021023, index sup 7% index, sibukiyaku sup 6% index, links sup 5% register, sibukiyaku sup 5% oracle02oct3, taikai24 sup 4% index, koukai sup 4% oracle02oct3, index sup 4% taikai17, index sup 4% Events, oracle02oct3 sup 3% Events, taikai24 sup 3% taikai20, index sup 3% taikai23, index sup 3% Events, taikai20 sup 3% Houjin, register sup 3% Houjin, sibukiyaku sup 3%
ac ドメイン 2item set	hit 66 index, register sup 19% Events, index sup 15% index, koukai sup 10% index, sibukiyaku sup 10% VLDB2001, index sup 9% register, sibukiyaku sup 9% VLDB2001, koukai sup 7% VLDB2001, register sup 7% kouenkai20020318, index sup 7% taikai19, index sup 7% taikai20, index sup 7% koukai, register sup 7% Dlib, index sup 6% SIGMOD, index sup 6% VLDB2001, Events sup 6% Events, koukai sup 6% taikai13, index sup 6% taikai18, index sup 6% index, mailing sup 6% Dlib, register sup 4% SIGMOD, VLDB2001 sup 4%	hit 126 taikai21, index sup 15% index, register sup 14% Events, index sup 9% taikai20, index sup 7% enquete, index sup 7% index, sibukiyaku sup 7% Events, taikai21 sup 6% register, sibukiyaku sup 5% taikai17, index sup 4% Dlib, index sup 3% taikai19, index sup 3% taikai20, taikai21 sup 3% taikai22, index sup 3% index, koukai sup 3% Dlib, register sup 3% VLDB2001, index sup 3% Events, taikai20 sup 3% taikai13, index sup 3% index, mailing sup 3% VLDB2001, taikai21 sup 2% Events, taikai19 sup 2%	hit 128 taikai23, index sup 20% index, register sup 17% index, koukai sup 15% index, sibukiyaku sup 13% Events, index sup 12% taikai22, index sup 11% index, links sup 11% VLDB2001, index sup 9% register, sibukiyaku sup 9% Dlib, index sup 7% VLDB2001, koukai sup 7% Events, koukai sup 7% taikai20, index sup 7% koukai, register sup 7% Events, taikai23 sup 6% taikai21, index sup 6% index, mailing sup 6% koukai, sibukiyaku sup 6% ICDE, index sup 5% SIGMOD, index sup 5% VLDB2001, taikai23 sup 5%	hit 217 Events, index sup 19% index, register sup 19% index, sibukiyaku sup 15% index, koukai sup 12% register, sibukiyaku sup 11% oracle02oct3, index sup 11% taikai24, index sup 10% datamining, index sup 10% Dlib, index sup 9% Houjin, index sup 9% VLDB2001, index sup 8% index, links sup 8% kouenkai20021023, index sup 8% Events, koukai sup 6% index, mailing sup 6% VLDB2001, koukai sup 6% Dlib, sibukiyaku sup 5% taikai23, index sup 5% mailing, register sup 5% mailing, sibukiyaku sup 5% Dlib, register sup 5%
com ドメイン 2item set	hit 484 No Itemset	hit 549 kouenkai20000526, mousikomi sup 2%	hit 780 Events, koukai sup 2% index, koukai sup 2% register, sibukiyaku sup 2%	hit 1040 Events, taikai23 sup 2% Events, Whoswho sup 2% oracle02oct3, Whoswho sup 2% Dlib, oracle02oct3 sup 2%

図 17　2002 年 other，co, com ドメインの四半期別高頻度アイテムセット

イニングを行うことができるアイテムセットキューブ機構の提案と実データへの適用について述べた．

3 節で述べた Cubic-Apriori とロールアップ処理を用いることで，アイテムセットキューブの実装を可能としている．特にロールアップ処理では，実体化をやり直す場合と比べて処理時間が 4 分の 1，使用メモリ量は 200 分の 1 に抑えることができるため，オンライン上のマルチユーザによる任意の問い合わせへの利用も考えることができる．

これにより，数値のデータキューブと実装されたアイテムセットキューブを併用することで，様々な条件で数値グラフを描き，グラフから発見した着目点での高頻度アイテムセットを即座に得て分析を行う，データマイニングが可能となった．

実データに対して，数値のデータキューブとアイテムセットキューブを利用したデータマイニングを行った結果，従来の数値による分析ではわからなかったイベント種類別ヒット数の増減の原因や，ドメインの特性 (人間であるか，巡回ロボットであるか等) を理解することができ，アイテムセットキューブ機構の有効性を示すことができた．

今後の課題は，[4] のような、多くの問い合わせに対応する必要のある大規模なログ分析用データセンターへの適用や，出力された高頻度アイテムセットを分析者の用途に応じた情報に変換するためのアプリケーションの作成がある．

文　　献

[1] Tadashi Ohmori, Yuichi Tsutatani, Mamoru Hoshi, A Novel Datacube Model Supporting Interactive Web-log Mining, IEEE First International Symposium on Cyber Worlds, pp.419-427, (2002)

[2] 助川，大森，星，蔦谷，Web ログ分析における高頻度アクセスパターン検出を支援するデータキューブモデル, DEWS2003, 1-A-01, (2003)

[3] 大森，成瀬，星，高頻度アイテムセットによる多次元的なログデータ分析を支援するデータキューブ機構, FIT2004, D-021, (2004)

[4] 1 日 4 億件のアラートをアナリストが分析する場所〜Symantec の SOC に潜入, http://internet.watch.impress.co.jp/cda/special/2004/03/23/2524.html

[5] 高田，見えログ：人間による計算機ログ解析を支援するログ情報ブラウザ, DBWeb 2003, 情処 DBS 研, pp.171-177, (2003)

[6] K L Wu, P.S. Yu, and A. Ballman, Speed Tracer. A Web Usage Mining and Analysis Tool, IBM Systems Jounal, Vol. 37,No 1, pp. 89-105, 1998

[7] 大塚,「Web mining」チュートリアル (Web usage mining), 電子情報通信学会 Web インテリジェンスとインタラクション研究会,Feb,2005

[8] analog, http://www.analog.cx/

[9] 石井，顧客管理における集合概念の重要性-航空会社を例として-, ACM SIGMOD 日本支部 第 30 回大会講演論文集，pp.37-49,2004

[10] インターネットマガジン 2004 年 3 月号, 株式会社インプレス, pp76-97

[11] R. Agrawal, et al.， Fast Algorithms for Mining Association Rules, Proc. 20th VLDB, pp.487-499, 1994.

[12] B.Prasetyo , et al.，Naviz:User Behavior Visualizations System using Web Access Log, 情報処理学会第 64 回全国大会 6X-01, 2002.

[13] R.Kohavi, et al, KDD cup 2000 organizers' report:peeling the onion, SIGKDD explorations vol.2(issue 2)，pp86-93，Dec，2000.

[14] S. Chaudhuri, U.Dayal, An Overview of Data Warehousing and OLAP Technology, SIGMOD Record, Vol. 26, No. 1, pp. 65-73, 1997.