1U-7

データマイニング向けデータキューブ機構によるログ分析の評価

成瀬正英　　大森匡　　星守　　山下由展

電気通信大学大学院情報システム学研究科*

1　概要

近年、ログデータを分析、理解する必要性は高まっている。著者らは、時間やイベントといった多属性で構成されているログデータマイニングを支援するデータキューブ機構として「アイテムセットキューブ」を提案している [1]、[2]。本稿では、アイテムセットキューブの実データへの適用とその評価について述べる。

2　アイテムセットキューブ

ログデータは、時間やユーザグループ、システム内で起きたイベントといった多属性から構成されている。分析手法の主流として、数値のデータキューブがある。この機構では、分割を与えた属性を次元としている多次元の空間に、該当するログ数を格納しており、問い合わせに対して、roll-up や slice といった演算によってキューブの形を変形させて、高速に数値を返すことができる (図 1)。このようにして、必要な数値を算出し、グラフを用いることで着目点を発見することができる。

しかし、数値による分析を行うと着目点ではどのような事象 (アイテム) の組合せがよく起きているかを調べたい要求がでる (図 1)。そのため、着目点での条件を満たすデータ群から、高頻度アイテムセットを算出する必要がある。だが、着目点を発見するたびにアイテムセットを計算していては、効率の良い分析を行うことができない。

そこで、著者らは、キューブ構造の各セルにその論理式を満たすデータ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機構「アイテムセットキューブ」を提案している。この機構では数値用データキューブと同じく、roll-up 等によりキューブを変形することで任意の問い合わせに対しても、高速にアイテムセットを返すことができる (図 2)。

3　評価

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。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 : 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、前出のいずれかに属する (all))、である。

図 3 は、数値のキューブを「月」-「イベント種類」の 2 次元にロールアップし、その内容をヒストグラムで表現したものである。この図からは、9 月〜12 月にかけて、all と講演会を見たグループが同じ様な増加の曲線を辿っていることがわかる。そこで、この時期に講演会を見たグループ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ため、アイテムセットキューブを同様にロールアップした (図 4)。図 4 からは、講演会を見たグループが 4Q(第 4 四半期) では講演会と組み合わせて、datamining や oracle02oct3 といった特別なイベントを見ていることがわかる。

次に、上記で着目した講演会を見たセッションレコード群の動作を見るためにドメイン種別の次元を追加する。数値のキューブをドリルダウンした結果 (図 5 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 5 下部である。図 5 下部から、other ドメインは図 4 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 月の講演会を見たグループの増加は、other ドメインが主要因であることがわかった。

本稿では、アイテムセットキューブの実データへの適用とその評価について述べた。実体化、ロールアップ等の演算高速化については、[1] で述べている。

参考文献.

1. 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)

2. 大森, 成瀬, 星, 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 (投稿中), DEWS2005,(2005)

*A new data cube system for multi-dimensional log data mining, M.Naruse, T.Ohmori, M.Hoshi, Y.Yamashita, U.Electro-Comm.

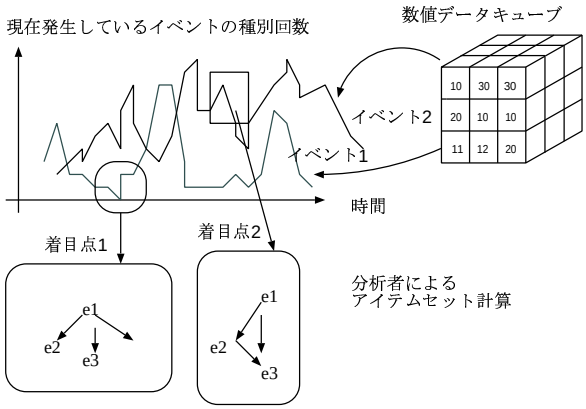


図 1: 数値グラフにもとづくアイテムセット計算の例

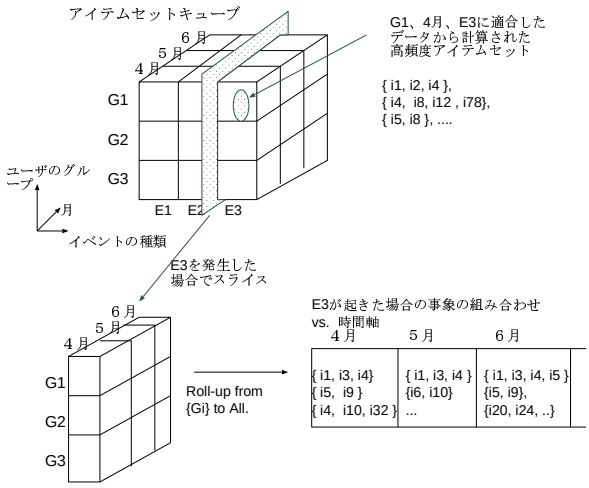


図 2: アイテムセットキューブとその用法

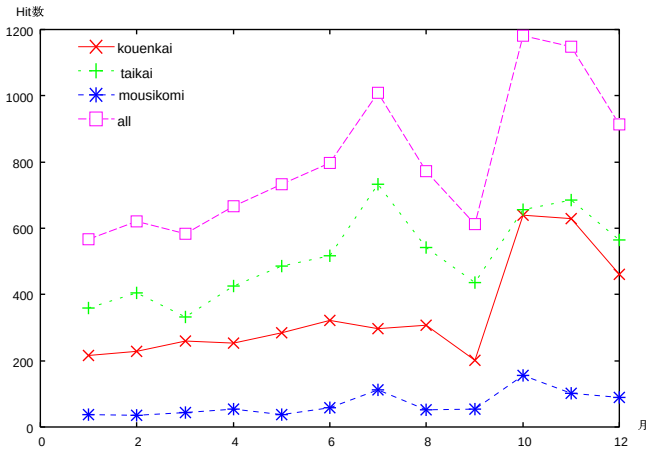


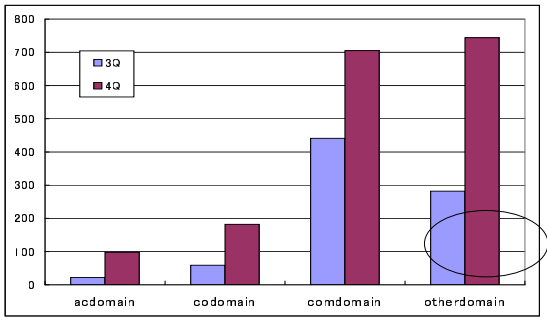
図 3: 各イベント種類の月別ヒット数

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 4: 講演会のいずれかを見たユーザが組合わせて見た他ページ (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 5: 講演会を含む、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