Webログ分析における対話的パターン検索のためのデータキューブモデル

6X−04

蔦谷 雄一　大森 匡　星守

電気通信大学大学院 情報システム学研究科*

1 研究の背景と目的

近年のDB分野では, World Wide Web等の情報サーバのアクセス記録データを大量に蓄積し, それらを分析して有用な情報を取り出す試みが盛んである. Webログを分析する際よくあるパターン抜き出す手法としてAprioriが用いられる[1]. しかしAprioriを用いてWebログを分析すると多くの情報が出てきてしまい, 理解するのは困難であるため, 分析条件を変えて対話的に分析を繰り返す必要がある[2].

そこで, 本稿ではAprioriを用いてWebログを対話的に分析するための分析モデルとしてデータキューブに基づいた分析モデルを提案し, スライス, ロールアップに相当する演算をサポートすることで, 対話的に分析を行なう方法とその演算の計算方法について述べる.

2 Webログ分析

Aprioriを用いてWebにおけるよくあるアクセスパターンを抜き出すためには, Webログをあるユーザーが一定時間の間にどのページを見たかというアクセスレコードの形に変換する必要がある. アクセスログからの変換は同一のIPaddressのログを時間を条件につなぎ合わせることで行なう.

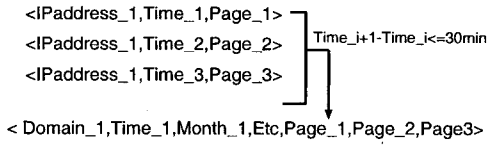

図 1: データ変換

本稿ではさらにTimeを時間と月に変換し, IPaddressをドメインに変換し, その他(Web上でのイベントの有る無し)を追加し(図1), 分析の対象として＜ドメイン,時間,月,その他,「ページ」＞の5属性の形のアクセスレコードを用いる. (但し, ここで言う属性「ページ」は図1の様にページ集合を値にとる).

このアクセスレコードから良くあるパターンを抜き出すのにAprioriを用いる. しかし, 単純にAprioriを実行しても多くのパターンが出てきてしまい, どの情報が有用なのか分からない. そこでApriori実行時に制約条件をつけ, 何回も対話的に分析を繰り返す必要がある.

*Data Cube Model for Interactive Mining of Web-log analysis. Y.TSUTATANI, T.OHMORI, M.HOSHI, Univ. Electro-Comm.

3 データキューブモデル

対話的に分析を繰り返すには, 出てくる情報の比較や, 制約条件の制御が簡単に行なえなければならない. そこで対話的に分析を繰り返すための分析モデルとしてデータキューブモデルを提案する(図2).

3.1 用語説明

本稿で扱うアクセスレコードは**属性**の組みである. 属性は, 原子的な値を取る場合と, (「ページ」のように)集合を値として取る場合がある. 一方, Aprioriでは, 命題変数を**アイテム**と呼び, アイテムから構成された(良く起きる)パターンを検出する. このパターン, すなわちアイテムの組合せを**アイテムセット**と呼び, 良く起きるパターンを**高頻度アイテムセット**と呼ぶ. 任意の1アイテムはあるカテゴリに所属するとされる. 本稿では, レコードの属性をこの(アイテムの)カテゴリに対応させて考える. 即ち, Webログの場合, ドメイン, 月, 「ページ」等が属性であり, あるドメイン d_i やあるページ p_j 等がアイテムになる(図2a). 結果, アクセスレコードはそのレコードで成立するアイテムの列で表される.

アクセスレコードにおいて属性 X のアイテムが, 必ず1つしか存在しないならば, 属性 X を**リレーション属性**と呼び, 1つ以上のアイテムが存在する可能性があるならば, 属性 X を**アイテムセット属性**と呼ぶ. Webログの場合, ドメイン, 時間, 月, その他がリレーション属性であり, 「ページ」がアイテムセット属性である.

3.2 定義

ある属性の組みからなる空間を考えた時, 各属性を軸として構成される空間を**キューブ**と呼ぶ. 図2(b)はA, B, Cを軸とした3次元のキューブである. キューブの桝目の一つ一つを**セル**と呼ぶ. セルは, 各属性のアイテム $a_i \in A, b_j \in B, c_k \in C$ が成立するレコード群から計算されたある値を持つ.

本稿では, **キューブ1**と**キューブ2**の2種類を用意する. キューブ1は, セルの値として, HIT数を持つ. **HIT数**とは, セルに対応するレコード群の総数のことである.

一方, キューブ2は, セルの値として, セル計算範囲のアクセスレコード群から求められた(アイテムセット属性「ページ」のアイテムからなる)高頻度アイテムセットの集合を持つ. これから, **Web巡回グラフ**(ユーザがWebをよくたどる経路)を計算できる. 以下では, キューブ2はセル値としてこのグラフを持つと見なし話を進める. また, キューブ1のセルの値はスカラーであるが, 軸

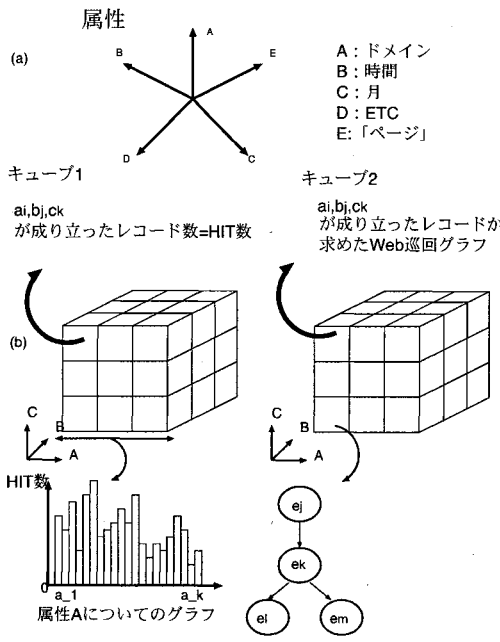

図 2: データキューブモデル

にアイテムセット属性が有る場合, 従来のデータキューブとは異り単純に集計演算等を行なうことが出来ない.

3.3 演算

キューブ 1, キューブ 2 に対して**スライス**および**ロールアップ**演算を適用することで効率的にユーザーの行動パターンの抽出と分類を行なうことができる.

・スライス

条件を指定して, 指定された条件を満たす各セルの値を計算する. キューブ 2 をスライスする時はスライスされたセルの値を相互に比較するために, 各セルの計算する情報の粒度を一定にする. (足切り値率 θ%を揃える).

・ロールアップ

ある属性のアイテムの幾つかを一つにまとめることでグループ化し, セルを新たに作り直す.

3.4 分析手順の例

キューブ 1 の値を見ること (図 3(a)) で大まかな分析の方針を立て, ヒット件数の多いドメインと少ないドメインにグループ化 (図 3(b)) して, それぞれのキューブ 2 の値の計算を行なってグループ毎の行動の違いを見たり, さらに月でスライスすることでそれらのグループの行動を時間軸に沿って見ることが出来る (図 3(c)). このようにスライスとロールアップを繰り返すことで図 3 の様な結果を得ることが可能である.

3.3 節で説明したスライスは, 図 3(a) や (c) を各々計算することである. ロールアップは図 3(b) の様にグループ G1, G2, G3 を作成し分析の階層を上げることである.

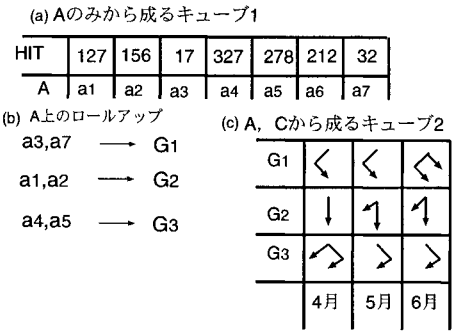

図 3: 分析結果一例

3.5 アルゴリズム

リレーション属性のみでスライスする場合, セルの値を計算する時の各セルに対応したレコードの一つ一つは重複しないが, アイテムセット属性でスライスする場合は重複する可能性があるため, それぞれ個別にセルの値を計算するとデータベースのスキャンに無駄が生じ, 候補アイテムの数え上げにも重複が生じる. そこでアイテムセット属性のアイテム p_i $(1 \leq i \leq N)$ に対応したセル i のキューブ 2 の値を足切り値率 θ% で計算する際のアルゴリズムを説明する.

[手順 1 候補アイテムセット生成]

あるアイテムセット I が候補アイテムセットであるためには I の部分集合の全てがあるセル i で高頻度アイテムセットでなければならない

[手順 2 候補アイテムセットの各セル毎のサポート数の数え上げ]

p_i $(1 \leq i \leq N)$ の内一つでも成り立つアクセスレコード群に対して, あるレコードで p_i が成り立つならば, セル i について候補アイテムセットのサポート数の数え上げる

[手順 3 高頻度アイテムセットの生成]

あるセル i で候補アイテムセット I が高頻度であるためには I のサポート数が $hit_i \times \theta/100$ 以上でなければならない (hit_i はキューブ 1 のセル i の値). 手順 1 に戻る.

4 まとめ

本稿では, 対話的な分析を行なう為の問題点を整理し, 解決の方法としてデータキューブモデルを提案し, さらにスライスの高速化のアルゴリズムを提案した.

今後の課題として得られた情報の統計的な手法への適用と, ロールアップの計算の高速化等を考えている.

参考文献

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

[2] U. Dayal, Web Content and Usage Mining for E-Commerce Applications, Proc. DBWEB2000, pp.311-316, 2000.