論文の内容の要旨

論文題目　Parallel Platform for Large Scale Web Usage Mining
（大規模ウェブログマイニングの為の
並列プラットフォームに関する研究）

氏　名　　イコ　プラムディオノ

本研究は、Webユーザの行動を記録した大規模のウェブアクセスサーバのアクセスログを解析できる並列ウェブマイニングプラットフォームの開発に関して、標準ハードウェアから成る大規模 PC クラスタ上での実装と性能評価を中心に研究した成果をまとめたものである.

近年、World　Wide　Web（以下Webと略する）の急速な発展でWeb全体の構造やコンテンツおよびWebを利用するユーザの行動を解析するウェブマイニングが注目されている．ウェブマイニングはウェブデータを解析するデータマイニング技術と定義されている．本研究では、主にユーザ行動を記録する大規模のアクセスログを対象にしている.

今日のウェブサイトは基本的には静的なWebページを用意しておき、全てのアクセスに対して同一のページを提供しているが、当該ユーザにとって必要な情報だけをそのユーザのコンテキストに合わせて、選別して提供する技術、すなわち、パーソナリゼーション技術が重要になってくる．さらに携帯電話をはじめとするモバイルデバイスからのインターネット利用が急激に増加しており、個別ユーザに合う先読み及びキャッシュも不可欠である.

そのためにウェブログ等に蓄積されているユーザ履歴から有用なアクセスパターンの発掘が必要不可欠だが，その処理は多くの計算量を要する．Yahoo は一日のユーザログは約40GBに上ると1999年5月のWWW8国際会議で発表している．ユーザのアクセスパターンを抽出するには少なくとも1カ月程度のログデータのマイニングが不可欠であるから、1TB以上の大容量データということになる.

そのようなタスクを処理できる並列プラットフォームとして，コストパフォーマンスの高いPCクラスタ上で，ウェブアクセスパターンマイニングシステムを提唱し，実際に構築した．システム設計で，データマイニングエンジンとしてPCクラスタを組み合わせることで柔軟性の向上，スケラビリティー及び応答時間短縮という目的が達成できる.

また、大規模なウェブサイトにはバックボーンデータベースからウェブページを生成する動的ページが主流になりつつあるが、メタデータによるCGIパラメータ抽出とデータベーススキーマ設計で動的ページのアクセスログにも対応できるようになっている.

プラットフォームの中心となるものはpattern-growthという現代最頻出パターンマイニングアルゴリズムのパラダイムに基づく並列アルゴリズムである．そのパラダイム

の代表的なFP-growthという頻出パターンマイニングアルゴリズムは、トランザクションデータベースをFP-treeというメモリ上データ構造に圧縮することで、それまで発表されたアルゴリズムよりはるかに高速であることが示された．しかしツリーような特殊なデータ構造の並列実行は、PCクラスタのような無共有型並列計算機では、特に困難である．本研究では、投入されるノード数に得られる高速化、いわゆる台数効果を向上させるために、並列実行単位の粒度を予測できるpath depthと呼ばれるパラメータを活用して、実行ノード間の負荷を均等化する新たなメカニズムを開発し、３２台の実行ノードで２３倍高速化されたことを図２（右）で示されている．図２（左）に従来の並列アルゴリズムであるHPAに比較した結果では一桁以上高速化されている．

さらにPCクラスタ上の実装では、ツリー構造がメモリ上に収めなければならないという制約のため、PCクラスタノードに分散したツリー構造に消費される全体メモリがシステムのスケラビリティーを決定する要因である．その全体メモリの最適化として、共通のアイテムヘッダーを持つツリー枝の動的な移送も提案した．またパソコン技術の急速な進歩により、PCクラスタを構成するノードを追加する際には新ノードが必ずしも同様の構成が得られるわけではなく、CPU 能力の異なるパソコンが混在するヘテロ環境になるのである．そのような環境下で、すべてのノードに従来のアルゴリズムを実行すると旧ノードがシステム全体のボトルネックになりかねず、異なる構成に応じる負荷分散が必要になる．提案されたメカニズムはヘテロ環境にも対応できることになっている．

頻出パターンマイニングに時系列という制約を加えたウェブアクセスパターンマイニングも同じフレームワークで実現できるが、従来の方式では、通信量が大きくなり、効率が悪化する．良好な台数効果を得るために通信圧縮の新しい方式も開発した．従来の方式では一つのユーザセッションで index.html が繰り返し訪れられるような複製アイテムに対して新たなサブデータベースを生成し、他ノードに送信するが、直接、ツリーのノードにあるアイテムのカウントに反映することにより、余分なサブデータベース生成を回避できた．

抽出されたウェブアクセスパターンの数が通常、膨大のため、人間が直感的に理解しやすいようにユーザ行動可視化ツール Naviz をプラットフォーム上実装した．Naviz はユーザのアクセスパターンの関連性をノードの配置パラメータとすることで複雑な解析結果を直感的に理解することができるだけではなく、様々な検索条件の下で個々ユーザアクセスパターンを追跡・比較することもできる．

プラットフォームの実応用としてNTT研究所のモバイルインフォサーチ（MIS）及び（株）NTT番号情報のi-Townpageサイトにおける実ウェブログを対象にしたマイニングを行った．I-Townpageログマイニング応用として問合わせ拡張による推薦システムも開発した．そのシステムはi-Townpageの業種階層構造を考慮してアクセスパターンのクラスタによる推薦を行っている．

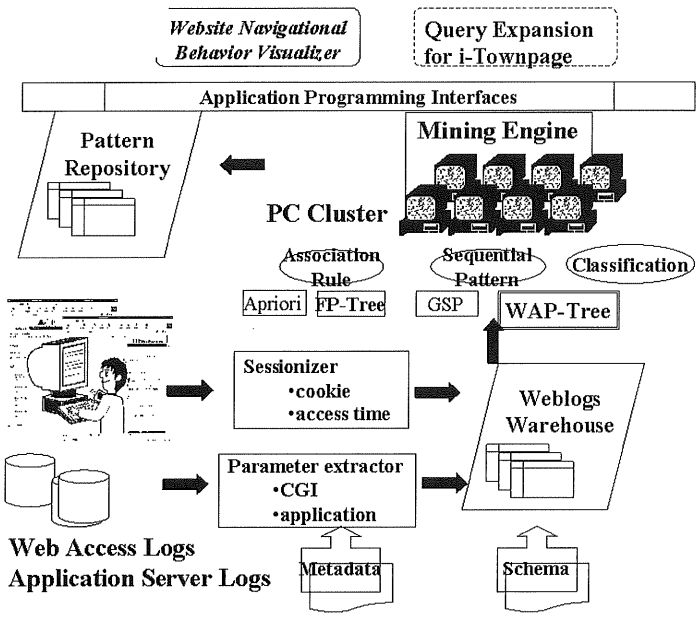

図1．ウェブマイニングシステム

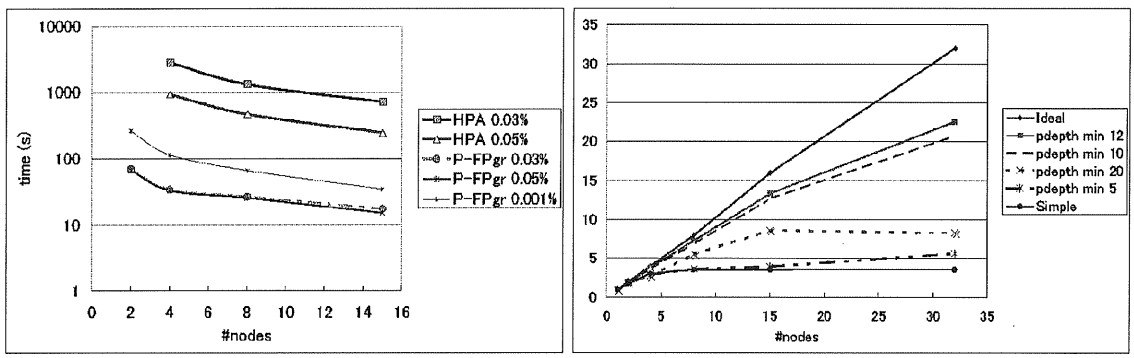

図2．HPAとの比較（左）　台数効果（右）