マルチメディア通信と分散処理 108－15
（２００２．６．７）

ウェブサイトのアクセス分布図を提供する視覚化システム

梶永 泰正　伊藤 貴之　山口 裕美　池端 裕子

日本アイ・ビー・エム(株) 東京基礎研究所
E-mail: kajinaga@trl.ibm.co.jp

ウェブサーバーのアクセスログファイルを入力データとして、そのアクセス分布図を自動作成し、ウェブで公開するシステムを提案する。本システムでは、アクセスログファイルに記録された URL のウェブページの階層構造を作成し、「データ宝石箱」という視覚化手法によってサイトマップを自動作成する。それと同時に、ユーザーが指定した属性を用いてアクセスを集計し、集計結果をサイトマップ上で表現する。筆者らは、以上の仕組みをウェブサーバー上で稼動し、視覚化結果を SVG (Scalable Vector Graphics) 形式ファイルで保存することにより、ウェブ上でアクセス分布図を公開できるようにシステムを構築中である。

A Visualization System That Provides the Website Access Distributions

Yasumasa KAJINAGA　Takayuki ITOH　Yumi YAMAGUCHI　Yuko IKEHATA

IBM Research, Tokyo Research Laboratory

Given the access log files from the Web server as input data, here we propose a visualization system that automatically generates the access distribution diagrams and publishes them on the Web. First, the system creates the hierarchical data structure of the Web pages from the URLs recorded on the access log files. Then it automatically generates the sitemap of those web pages by what we called, “the data jewel box” visualization method. In parallel, it counts access numbers as to the user-specified attributes, and represents the statistics over the sitemap. Authors are currently constructing such a system that runs on the server, saves the analyzed access data into SVG (Scalable Vector Graphics) format, and visualizes access distributions publicly available on the Web.

1. はじめに

インターネット上で情報を共有するシステムである World Wide Web (WWW：以下簡単のためウェブと呼ぶ)が誕生して、既に 10 年以上が経過した。この間、様々な技術や応用が作り出され、ウェブは単なる情報共有システムに留まらず、様々なサービスを生み出す社会的インフラストラクチャーとして機能し始めている。

それに伴って、インターネットを利用する人口も増え、非営利的・営利的の区別を問わず、ウェブ上でのサービスを利用も、つい数年前では考えられないような数にまで増えている。

例えば、筆者らの勤務する日本アイ・ビー・エムのウェブサイトでは一日に 10 万件近くのユーザーアクセスがある。

また Yahoo のような巨大なポータルサイトにおいても、オークションのサービスは何万件もの取引が登録されている事も良く知られているだろう。

またこのようなユーザー数の増加と並んでサービスを提供する側も拡大しており、2000 年の時点で、インターネットに接続しているホストマシン

の台数は一億台に近づこうとしている[Reki]。

しかしながら、インターネットの発展は試行錯誤の連続ともいえる歴史的側面があり、ウェブ上でのサービスをいかに上手く行うかについて、体系化された手法や方法論といったものはまだ確立されていないといって良いだろう。

そのようなインターネットの科学ともいえるものを生み出すためにも、ウェブサーバー上に残されたログを解析し、ユーザー行動や望ましいウェブの運用やデザインについて知ることが、サーバーの管理者やコンテンツを制作する人たちにとって必要とされている[IM00][dW1][dW2]。

本研究では、ウェブサーバーのアクセスログファイルを入力データとして、そのアクセス分布図を自動作成し、ウェブで公開するシステムを提案する。また、そのために現在プロトタイプシステムを構築しており、それについて詳しく述べたい。

ウェブサーバーのログからユーザーアクセス履歴などを集計するシステムは、フリーウェアとして流通するものから商用の製品まですでにいくつか存在している。筆者らが知る限り、これらの多くはログファイルをバッチ処理で解析し、リスト形式で上位何十かのアクセスについて HTML ファイルを出力したりするものから、アプリケーションサービスプロバイダとして契約したサイトからデータを預かり解析したと形態は様々である。いずれの場合も、解析結果をテキストと簡単なグラフで表示する静的なインターフェースであり、サイト全体での傾向を概観することや、対話的な形でログデータの中を見て回るような事は実現されていない。

本研究では、アクセスログファイルに記録された URL のウェブページの階層構造を作成し、「データ宝石箱」という視覚化手法[Ito01]によってサイトマップを自動作成する。それと同時に、ユーザーが指定した属性を用いてアクセスを集計し、集計結果をサイトマップ上で表現する。さらに、以上の仕組みをウェブサーバー上で稼動し、視覚化結果を SVG 形式ファイルで保存することにより、ウェブ上でアクセス分布図を公開できるようにシステムについて述べる。

2. データとシステム構成

本章では本システムで用いるデータ構造とシステム構成について述べる。

2.1 標準アクセスログと内部データ構造

本研究で用いるウェブサーバー・ログは標準的に使われている NCSA ログ形式にものを用いた[AL1][AL2]。

表 1 標準形式のウェブサーバー・ログの内容

フィールド	内容	例
remotehost	クライアントの IP アドレスまたはドメイン名	123.4.56.78 webvis.trl.ibm.com
logname	クライアントの識別子	-
username	クライアント認証のためのユーザー名（ID）	dsmith
date	日付	[13/May/2002:08:45:32 +0500]
request	HTTP 要求	“GET /index.html HTTP/1.0”
status	HTTP 要求の成功失敗を表す数値コード	200
bytes	転送されたデータのバイト数（HTTP ヘッダーを除く）	1043
referrer	クライアントが移動してきた先の URL	http://www.ibm.com/ index.html
resource	クライアントが参照中のサーバー資源 URL	/jp/pc/banner.gif

次に、本システムで用いた中間データ形式について述べる。

今回はデータウエアハウスなどでよく用いられるスタースキーマ形式にまとめた[Cha97]。

各データは、SessionFacts というセッション(ユーザーがあるサイトにアクセスして別のサイトにアクセスするまでのアクセス履歴の一連)と HitFacts というウェブページなどへのアクセスに関する情報の二つのテーブルに分けられる。この二つはユニークな ID(primary key)によって相互に結び付けられている(図 1)。

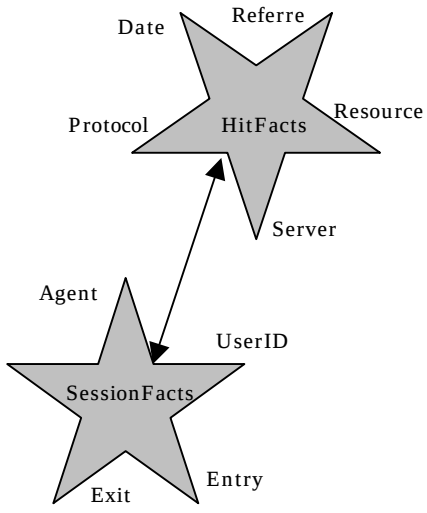

図 1 今回用いたスタースキーマ形式

この二つのテーブルはそれぞれに固有の情報と

個々の項目に関するテーブルへの参照ＩＤを含んでいる。ユーザーのアクセス状況を知るために javascript などを用いた高度な log 収集を行えば、数十項目のデータを集めることも可能だが、ウェブサーバーで得られるログは限られているため、項目はそれ程多くは無い。そこで、今回使用したなかで､しかも主要な項目のみを表 2 にまとめた。バッチ処理の都合上、これらを 14 個の csv ファイル(Access csv ファイル)に出力した。

表 2 HitFacts（上）と SessionFacts（下）ファイルの主要な項目

項目(HitFacts)	内容
HITS	各ヒット数(=1)
TIMETAKEN	サーバーがヒット処理に要した時間
HIT_ID	あるヒットに固有の ID
SESSION_ID	SessionFacts ファイル上のあるセッションへの ID
LOCALDATE_ID	Calender ファイル(年月日情報)上での ID
LOCALTIMEOFDAY_ID	TimeOfDay ファイル(時間情報)上での ID
RESOURCE_ID	Resource ファイル(サーバー資源 URL 等の情報)上での ID
REFERRER_ID	Referrer ファイル(クライアントが移動してきた先の URL 等の情報)上での ID
PROTOCOL_ID	Protocal ファイル(サーバーが使用した通信プロトコル)上での ID
REFPROTOCOL_ID	Protocal ファイル上(クライアントが使用した通信プロトコル)での ID
HTTPVERSION_ID	HttpVersion ファイル(HTTP のバージョン情報)上での ID
RETURNCODE_ID	ReturnCode ファイル(HTTP status code 情報、例えばエラーなど)上での ID
STATUS_ID	ResetStatus ファイル(network connection の status 情報)上での ID
JS_ID	JavaScriptStatus ファイル(クライアントの javascript 許可情報)上での ID
COOKIESTATUS_ID	CookieStatus ファイル(クライアントの coockie 許可情報)上での ID

項目(SessionFacts)	内容
SESSIONS	各セッション数(=1)
HITS	あるセッション中のヒット数
DURATION	あるセッションに掛かった時間
SESSION_ID	あるセッションに固有の ID
USER_ID	User ファイル(ユーザー情報)上での ID
REFERRER_ID	Referrer ファイル(クライアント URL 等の情報)上での ID
LOCALDATE_ID	Calender ファイル(年月日情報)上での ID
LOCALTIMEOFDAY_ID	TimeOfDay ファイル(時間情報)上での ID
USERAGENT_ID	UserAgent ファイル(クライアント web browser の情報)上での ID
ENTRYRESOURCE_ID	Resource ファイル(ユーザーが最初にアクセスしたサーバー資源の情報)上での ID
NETWORK_ID	Network ファイル(クライアントの IP アドレス情報)上での ID
EXITRESOURCE_ID	Resource ファイル(ユーザーが最後にアクセスしたサーバー資源の情報)上での ID

最後にこれらを解析してグループ化することで、net ファイルと呼ぶ独自形式のデータファイルにまとめる。これらは、前述の csv ファイルからサイト内のウェブページ URL を抜き出し、そこからサーバ上でのファイル階層を見ることで各ウェブページを階層化したものである。net ファイル中の主要なフィールドとその値を表 3 に示す。

表 3 Net ファイルの主な内容 いずれも固有の ID 番号が振られており、それぞれ総数を表すフィールド(num*field*)も存在する

項目	内容	例
calender	ID とアクセス日時(YYYY-MM-DD)	1 2002-5-13
timeofday	ID とアクセス時間(H:M:S)	2 1:4:16
resource	アクセスされたサーバー上の URL	3 /jp/pc/index.html
referrerhost	ユーザーがアクセスしてきた先のホスト名	45 www.trl.ibm.co.jp
referrerurl	ユーザーがアクセスしてきた先のホスト上の URL	678 /news/020513.html
returncode	http サーバーのリターンコード	1 200 Successful - Okay
network	クライアントの IP アドレス	234 123.45.67.89
nodehit	ある一つのヒットの固有 ID	12
nodelocaldate	あるヒットについての calender ID への参照	12 1
nodelocaltimeofday	あるヒットについての timeofday ID への参照	12 2
noderesource	あるヒットについての resource ID への参照	12 3
nodereferrerhost	あるヒットについての referrerhost ID への参照	12 45
nodereferrerurl	あるヒットについ	12 678

	ての referrerurl ID への参照	
nodereturncode	あるヒットについての returncode ID への参照	12 1
nodenetwork	あるヒットについての network ID	12 234

2.2 プロトタイプシステムの構成

ウェブのアクセス分布を可視化するシステムは、ウェブサーバー（WS）、アクセス分布の可視化エンジン（VE）、およびグラフィカルユーザーインターフェース（GUI）のように三層に分けられるだろう。実際の実装には Java (JDK 1.3)を用いた。クライアントサイドとも呼べるユーザーインターフェース（UI）部分だが、プロトタイプ（評価用）システムでは OpenGL の java-bindings の一つである GL4Java API を使った専用 GUI アプリケーションで結果を表示している（本稿の結果画像もそのアプリケーションの出力である）。この UI 部分は最終的には出力を SVG (Scalable Vector Graphics) 形式にすることで、ウェブブラウザ等で簡便に表示できるように実装される予定である。研究用プロトタイプの実行環境としては IBM Intellistation M-Pro (Pentium4、2.0GHz)上で実験を行った。

今回は特にこの VE 部分について詳しく述べたい。本研究では、前述のデータを処理するために VE 部分を次のようなにシステムを三層化した。

(a) Log file converter (LFC)：標準形式の access log file を本研究で採用したスタースキーマ形式の CSV ファイル群に分割する。
(b) Net file generator (NFG): スタースキーマ形式の CSV ファイル群から URL による階層構造を生成し、内部形式により構造と属性を記述した NET ファイル形式に変換する。
(c) Layout engine (LE): (b) の NET ファイルの階層構造から、サイトマップの座標値を計算する。データ宝石箱(後述)アルゴリズムを用いることで高速に座標値を決定する。

このように分割することで、(a)(b)部分の重い処理がバッチ処理化できる。詳しくは後述するが、一般のワークステーションレベルでは(a)や(b)でのテキストファイル処理はアクセスログが数万件に及ぶ大規模なサイトの場合数時間を要し、すべてを対話的時間内に処理するのは難しい。

中間部分でそれぞれファイルに出力しているが、これを DB 化することも可能である。それについては最後に述べる。

なお、この原稿を執筆中の現在は、(b)および(c)部分はモジュールとしては分離されているものの、プロトタイプのため、まだサーバー・クライアントシステムとして実装はされていない事はお断りしておく。

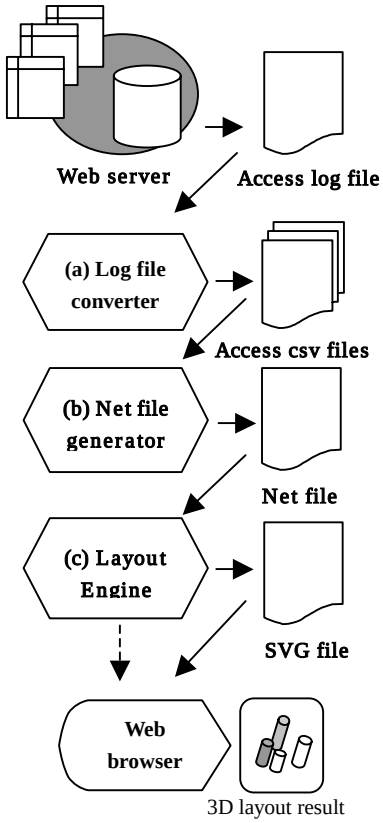

図 2 システム構成図 プロトタイプシステムでは SVG ファイルと web browser を用いず、独自形式のデータフォーマットファイルと専用 GUI アプリケーションで実験した

3. アクセス分布の視覚化インターフェース

ここでは前節で触れたユーザーインターフェース部分について述べたい。

3.1 データ宝石箱 －サイトマップの自動生成－

サイトマップの自動生成は、我々のグループで開発した「データ宝石箱」レイアウトを用いた[Ito01]。これは階層構造を入れ子型に表現するもので、計算幾何のアルゴリズム(逐次的 Delaunay 三角形メッシュ生成アルゴリズム)を用いて高速な配置を実現している。

図 2 にあるように、個々の点(塗りつぶしてある矩形のドット)が一つ一つのウェブページを表す。拡大するとサムネイル表示が可能である。ディレクトリは矩形の枠で表現され、階層構造は入れ子の構造として表現されている。

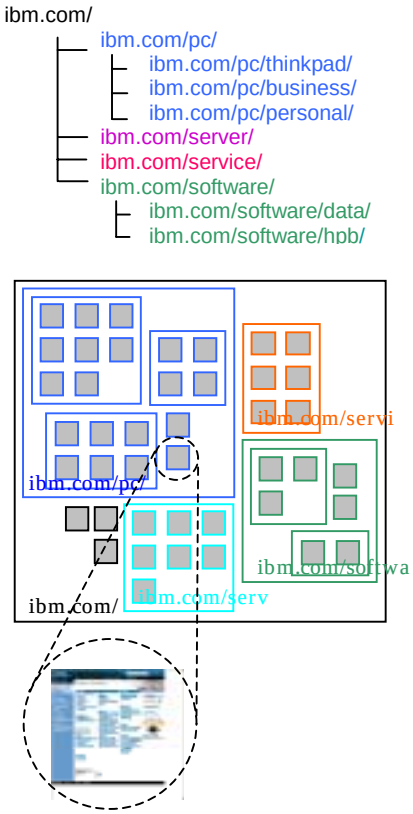

図 2 データ宝石箱レイアウト 計算幾何アルゴリズムを用い入れ子状の配置を高速に実現しサイトマップを生成する

3.2 アクセスグラフとサイトマップの連携

ここでは、次に述べるようにサイトマップと統計表示を連動させることでアクセル履歴の分かりやすい表示を実現しようとした[Yam02a]。

まず、前節で生成したサイトマップの”z 軸”方向に各ウェブページのアクセス数を表すグラフを付け、この表示を三次元化する。こうすることで、ウェブページ群でのアクセス分布が一目で見て分かる。

次に、統計表示のための棒グラフ表示部分について述べる。これを以下、アクセスグラフと呼ぶ。

図 3 にあるように、アクセスデータの統計的表示のために棒グラフを用いている。横軸はアクセスログデータの中からユーザーが関心のある項目について、縦軸はアクセス数である。横軸の項目としては、前章で説明したように day、time of the day、referrer、resource などが選択できる。また、グラフのひとつの棒(縦軸方向)をユーザーの次に関心のある項目で色分けして区分することも可能である。例えば、数日間のまとまったデータで、横軸に day 属性を選び、縦の区分に time of the day の属性を選べば、ある日の、ある時間帯のアクセス数が見てわかる。

次に、このアクセスグラフと前節で述べたサイトマップとの連携を述べる。連携の仕方は、アクセスグラフからサイトマップへ、またサイトマップからアクセスグラフへの二通りが考えられる。

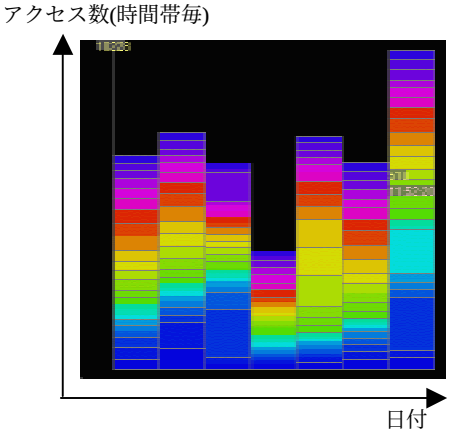

図 3 アクセスグラフ この例では日付毎のアクセス数を棒グラフ化しさらにそれぞれの日を時間帯に分けて色分けしている

[連携 1] アクセスグラフ中でユーザーの関心のある部分(ひとつの棒グラフ、もしくはその中の区分)を選択すると、サイトマップ上で、その部分に関連のあるアクセスがあったウェブページ上に前節で説明したようにアクセス数に比例した高さで棒を表示する。こうすることで、アクセスログのある部分に対応するアクセスがサイト全体にどのように分布しているかを読み取れる。

[連携 2] サイトマップ上でユーザーの関心のあるウェブページをアクセスすると、あらかじめ選択されていた属性により分類された、そのウェブページに対応するアクセスの統計データがアクセスグラフに表示される。

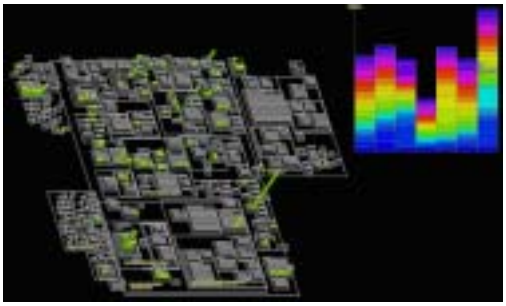

図 4 サイトマップとアクセスグラフの連携

前者はサイト全体の傾向を掴む事ができるし後者は個々のウェブページの(選択した属性値に従った)傾向を見ることが可能である。筆者らの実装結果を図 4 に示す。左側はサイトマップ、右側はアクセスグラフである。

4. プロトタイプシステムでの実験

ここではプロトタイプシステムを用いて実際のアクセスログデータを処理した時の実験結果を簡単に報告する。データはあるサイトの一週間分のアクセスログである。尚、紙面の都合上、結果を図示できなかったが、興味のある読者は文献[Yam02b]を参照して欲しい。

[ケース 1] アクセスグラフである日の朝にアクセスが集中している時間帯が見つかった。これを[連携 1]によりサイトマップ上で見てみると特定のページへのアクセスが集中していた。さらに[連携 2]を用いて参照元を見てみるとある新聞社サイトからである事が分かった。実はそのサイトのオンラインニュースにアクセスされているページ内容の紹介とリンクがあった事がわかった。

[ケース 2] アクセスグラフのある時間帯のアクセスを[連携 1]により見てみると同一ディレクトリのほぼすべてのファイルがアクセスされている。ここで[連携2]を用いてユーザーのIPアドレスを見てみるとほぼ同一のクライアントからのアクセスであった。あるディレクトリのページ全てを閲覧するユーザーの存在が分かった。

この他、サーチエンジンロボットによる全面的なアクセスなども容易に見つけ出す事が出来た(この場合、アクセスがある時間帯に集中しているものの、各ページへのアクセス数は低く、テキストベースのランキングでは見つけにくいと思われる)。

5. まとめ

以上でウェブサイトのアクセス分布の視覚化を提供するシステムと、そのプロトタイプ実装と実験結果について述べた。本章ではシステムの利点や問題点等についてまとめたい。

まず利点について述べたい。従来手法の多くはテキストベースであり、統計結果の表示もおのおのの項目に対する比較的単純なものが多い。しかしながら、ウェブアクセスは多様なユーザーの行動が現れたものであり、統計要素を横断的に見る必要があるだろう。従来ツールでは、このような作業は経験と熟練を必要とするが、４章での実験結果でも分かるように、今回試したような可視化システムは容易に現象の理解を導く事が可能ではないだろうか。

実際、情報可視化の世界での評価法を視覚化インターフェースに適用した場合の結果や、筆者らの所属する会社のコンテンツをデザイン・管理する部門とのミーティングやアンケートでは概ね良好であった[Yam02b]。

問題点としては何よりシステムのパフォーマンスとしての評価が難しい点があげられる。一つには、サーバー上のデータを集めることは(企業などでは特に)難しい点、評価基準の設定が難しい点などが挙げられる。というのも、システム全体としてのパフォーマンス（アクセスログファイルの効率的な処理と、処理結果からのインタラクティブなデータ抽出）や、結果の視覚化部分のグラフィックス性能、さらにはユーザーインターフェースとしての使い勝手など多岐にわたり、評価者もサーバー管理者からコンテンツをデザインする人まで様々な人が存在する点にある。残念ながら、それらのすべてを含めての評価というのは時間的にもまだ出来ていないのが実情である。

筆者らは、このような研究が端緒となり、今後、このようなさらに研究が進み、そこでサービスを行おうとする様々な人たちに、あたかも PC 上でシステムのプロセスやパフォーマンスを見るように、あるウェブサイト上でのアクセスの様子が、対話的で柔軟に理解できるシステムが開発され、今後のウェブやインターネットの発展に役立てられる事を願っている。

謝辞

多くの助言と協力を下さった、日本アイ・ビー・エム(株)東京基礎研究所日高一義氏、梶谷浩一氏、青野雅樹氏、井上恵介氏、山田敦氏、土井淳氏に感謝の意を表する。

参考文献

[Reki]http://terakoya.yomiuri.co.jp/rekishi/index.html
[IM00] INTERNET magazine 2001/5 pp.188-207.
[dW1]http://www-6.ibm.com/jp/developerworks/web/010907/j_wa-mwt1.html
[dW2]http://www-6.ibm.com/jp/developerworks/web/010914/j_wa-mwt2.html
[AL1]http://httpd.apache.org/docs-2.0/logs.html
[AL2]http://httpd.apache.org/docs/mod/mod_log_config.html
[Cha97] S. Chaudhuri and U. Dayal, “An overview of data warehousing and OLAP technology.” SIMGMOD Record, 26(1), pp.65-74, 1997
[Ito01] 伊藤貴之、梶永泰正、池端裕子「データ宝石箱：大規模階層型データのグラフィックスショーケース」情報処理学会グラフィックスと CAD 研究会 104-15
[Yam02a] 山口裕美、池端裕子、伊藤貴之、梶永泰正「データ宝石箱を用いたウェブアクセスログの視覚化」可視化情報学会（投稿中）
[Yam02b] 山口裕美、伊藤貴之、池端裕子、梶永泰正「サイトマップ表示とアクセス統計表示の連携によるウェブサイト視覚化ツール」情報処理学会論文誌（投稿中）