3P-1

Web における情報配信の最適化のためのユーザ行動の分析手法の提案

向井 康人 ‡　大囿 忠親 ‡　伊藤 孝行 ‡　新谷 虎松 ‡
名古屋工業大学大学院工学研究科情報工学専攻 ‡
e-mail: {mukai, ozono, itota, tora}@ics.nitech.ac.jp

1　はじめに

WWW は個人から企業に至るまで広く情報配信の場として利用されるようになった．しかし，配信されたコンテンツが実際どのようにして閲覧され，実際に閲覧されたページが閲覧者にとって有効かどうかの調査を行うことは困難である．配信している内容やユーザビリティが適切であるかどうかということは，情報配信者にとって更新・修正の面で重要である．上述の問題に対して [1] のような Web アクセス解析ツールが存在する．上記解析ツールは Web サーバに蓄積されたアクセスログを解析することで，データをパラメータやグラフ化し，その結果サイトの閲覧や利用状況を把握することが可能となる．しかし，アクセスログにおけるデータは，閲覧されたページをユーザが有効であると感じるかどうかは関係なく，実際にアクセスされたという事実のみのデータであるという問題がある．そこで，我々は Web ページの内容やページ内でユーザが行った行動などを収集して分析することで，最終的にページがユーザに及ぼした影響の度合いを得られるのではないかと考えた．本稿では，ユーザの行動をデータとして収集し，収集したデータにルールを適応することで分析を行うことを可能にする一連のシステムについて述べる．

2　システム構成

本システムは Web アプリケーションとして実装される．本システムは，CGI，JAVA サーバ，スクリプト，およびデータベースから構築される．CGI はユーザへのデータを管理するためのインタフェースの役割を行う．図 1 に本システムのインタフェースを示す．図 1 ではユーザが管理するために登録しているページデータの一覧が表示されている．JAVA サーバはスクリプトの制御を行う．スクリプトはデータ収集などの機能を実現する．データベースはページ，コンテンツ，ルール，および収集するデータを格納するために設置される．

図 1: システムインタフェース：ページデータ一覧

る．本システムの流れは，まず，管理するためのページに関するメタデータを作成する．メタデータを作成することでページを管理するために必要なデータベースを作成する．作成されたデータベースは ID で関連づけられ，収集されるデータが格納されるデータベースは一意に定まる．データを収集すると同時にコンテンツ，ルールのデータを作成することで収集したデータからパターンを取り出す．

3　閲覧状況収集手法

本研究では，Web ページへの訪問者が実際にサイト上で行った行動をもとに分析を行う．サイト上で行った行動とは，例えば，マウスクリックや，スクロールのようなイベントが発生する行動のことを指す．また，上記の行動の分析において Web ページ上のコンテンツの情報が必要不可欠になる．何故なら，クリックをしたという行動があったとしてもそのデータだけでは何をクリックしたのか分析することは不可能であるからである．以上より本システムは，上記の行動と行動が行われた Web ページ上のコンテンツの情報を収集する．収集機能は Javascript により実装され，クライアントのバックグラウンドで動作する．ユーザを特定するためにスクリプトを書き出すときに JAVA サーバが一意に定まるユーザ ID を付加することでセッションの行動を特定できるようにしておく．

また，同時にコンテンツの大きさや位置などの情報を収集する．しかし，[2] にみられるように，HTML ベースのレイアウト記述言語で記述された Web 上の情報 (コンテンツ) は，計算機で直接扱うのは困難である

The Analysis of Users Bihavior for optimization of transmission of information on web

Yasuto MUKAI, Tadachika OZONO, Takayuki ITO, and Toramatsu SHINTANI
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, 466-8555 JAPAN

といえる．そこで，本システムではコンテンツに意味付けを行うために，前節で述べたコンテンツデータを作成する．本システムには，作成したコンテンツデータを収集したデータに付加するための支援機能をがある．本機能は Web ブラウザでページを開き，コンテンツデータを付加したいタグをクリックして，コンテンツデータを選択することで < タグ access_log=’ コンテンツデータ ID’> といった形に整形する．access_log パラメータにより，収集したコンテンツのデータと作成されたコンテンツ意味を関連づけることが可能となる．

以上により，Web ページ，Web ページ上でのユーザの行動，およびコンテンツの意味をコンピュータが認識できるという意味で関連づけることが可能になった．

4 ユーザ行動抽出ルールと分析

収集したユーザ行動はただの連続データである．この連続データは，ユーザ ID や対象ページ，マウスの座標やスクロールの位置などの情報で構成される．また，このデータからユーザの行動を分割することは困難である．そこで，本システムではルールをいくつか作成し，このルールを収集したデータに適応させることでそのルールに適合したパターンを抽出するという仕組みで成り立っている．以下にルール例を挙げる．(1) 連続したスクロールイベント (2) スクロール時間と移動距離を指定．このルールによって，一定間隔でスクロールしているパターンの抽出が可能となる．このルールと取得したユーザ行動データにより，Web ページ上におけるユーザ行動のいくつかのパターンを発見することができる．例えば，ユーザのスクロールの仕方に対していくつかのパターンをみることが可能となる．ユーザは常に同じ間隔でスクロールしているわけではないし，複数のスクロールの間にはマウスアクションなどが入り交じることが見て取れる． このようにして，ルールを適用させることでユーザ行動にパターンを発見し意味を持たせて分析することが可能になる．

5 評価

閲覧状況収集手法が実運用に適しているかどうかの評価を行う．評価としてはスクリプトを埋め込んだページ (以下ページ a) と埋め込んでいないそのままのページ (以下ページ b) においてページの上端から下端までスクロールするのにかかる時間の比較を行う．ページ a はイベント毎にバックグラウンドでデータ取得・送信処理をしているために処理に遅延が生じる．図 2 に実験対象としたひとつのページに対する結果グラフである．標準とはページ b のことを指し，改良後とはページ a のことを指し，差分とはページ a とページ b の遅

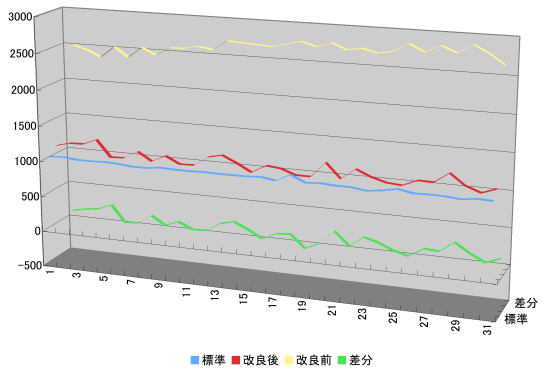

図 2: 比較実験の折れ線グラフ

延を指す．他のページでも比較的同様の結果が得られた．以下にこれらの結果から考察を述べる．ページ a とページ b のスクロール時間の平均の差分は 99 ミリ秒である．つまり，上端から下端にスクロールするときにかかる遅延は 0.099 秒である．また，このとき取得したデータ数は 8 であり，これはページ上で 8 回イベントが発生したことを表す．遅延はイベント処理が原因で発生することから，1 イベントにおける遅延は 99/8=12.375 ミリ秒である．同様にその他のページで測ったところページが含有するタグが多いと 1 イベントにおける遅延がおおきくなるものの 50 ミリ秒以内に収まっていたことがわかった．この遅延は非常に小さいことから本手法は実運用に適しているといえると考える．

6 おわりに

本論文では，Web ページ上でのユーザの行動とコンテンツを収集し，ルールを適応することでユーザ行動からパターンを抽出することができるシステムについて述べた．また，本収集手法が実運用に適していることを示した．本システムは，実際に Web ページ上でユーザがどのような行動をしているかという分析をする上で有用であり，ページの評価を行うためのシステムの有用な基盤であると考える．今後の課題は，ルールの多様化・作成支援を可能にすることで分析の幅が広がると考える．また，ルールから抽出されたパターンに得点を割り当てることで，分析対象ページをアクセス数のみに依存せずに効果的に評価することが期待できる．

参考文献

[1] GoogleAnalytics:http://www.google.com/analytics/

[2] 南野朋之, 齋藤豪, 奥村学:“ 繰り返し構造を用いた Web ページの構造化に関する研究 ”, 情報処理学会研究報告,2003-NL-154,pp.185-192,2003.