マウスカーソルの軌跡による Web 閲覧者の行動分析手法の提案 ～K 平均法によるクラスタリングを用いた場合～

A Proposal of Web Browser's Behavior Analysis Method by Mouse Cursor Track When using clustering by K-means clustering

椋本大輔 †, 森山真光 †
Daisuke Mukumoto†, and Masamitsu Moriyama†
† 近畿大学大学院　総合理工学研究科
†Graduate School of Science and Engineering Research, Kindai Univ.

要旨
近年, Web サイトによる情報発信は重要な位置を占めている. Web 閲覧者の動向を調査する上で有効な手法にマウスカーソルの軌跡を用いた分析がある. 既存の手法では, 時間毎, 空間毎に分析する手法が多く存在するが, 膨大な軌跡から Web 閲覧者毎の特徴を掴むのは困難である. そこで本研究では, マウスカーソルの軌跡を IP アドレス毎に分類し, K 平均法を用いたクラスタリングによる行動分析手法を提案する.

1. はじめに

近年, インターネットやスマートフォンの普及により個人から企業まで幅広く Web サイトを運用, 管理をしている. 企業対企業での取引を Business to Business (以下 BtoB), 企業対個人の間で行われる取引を Business to Custormer (以下 BtoC) と呼び, 情報収集方法, 情報発信方法としても Web サイトの利用割合は高いものとなっている. そのため企業が Web サイトに取り組む事は重要な位置を占めており, アクセス回数の増加, Web 閲覧者の媒体に合わせた Web サイトのデザイン変更などを行う事が必要になっている. その方法の一つして Web サイトのアクセスログを用いたアクセスログ解析が挙げられる. 一般的なアクセスログには, リクエスト時刻や URL などクライアントからサーバへアクセスが起こった際のデータが蓄積されており, これらから Web 閲覧者のページ遷移や検索エンジンからどの様な単語でその Web サイトに行き着いたのかを分析する研究が行われている. また Web 閲覧者が Web サイト内でどのページに興味を示したのか, どのコンテンツに興味を持ったのかといった Web 閲覧者の動向を調査する方法としてマウスカーソルの動きを捉えたマウスログから軌跡として可視化, 分析する手法も存在する [1][2]. この手法によってページ上の座標といった空間的要素毎及びアクセス時間といった時間的要素毎にマウスログを可視化することは可能になるが, それでも膨大なマウスカーソルの軌跡から Web 閲覧者毎の特徴を掴むのは困難である.

そこで本研究では, Web ビーコン型のマウスログ取得スクリプトを用いることで得られたマウスログから, アクセスされたページに対し Web 閲覧者のマウスカーソルの軌跡を空間的要素と時間的要素毎に可視化する手法に加え, Web 閲覧者がどの部分に注目するのかを知るためにマウスログからマウスカーソルの軌跡を K 平均法を用いてクラスタリングし, クラスタ毎に分類, 可視化する行動分析手法を提案する.

2. マウスログの取得方法とマウスカーソルの軌跡の可視化

本稿ではマウス動作の傾向を調べるためにマウスカーソルの動きを捉えたマウスログを用いる. マウスログの取得には JavaScript と Web サーバ を用いた Web ビーコン型マウスログ取得方法を用いる. 図 1 に使用する Web ビーコン型のマウスログ取得スクリプトによって送信されるマウスログのフォーマットを示す. フォーマットは Web 閲覧者の IP アドレスやアクセスされた時刻 (Unix time) などの情報に加え, マウスカーソルの X 座標, Y 座標や使用している画面のサイズを JSON 形式で纏める.

図 2 に Web ビーコン型のマウスログ取得スクリプトの概要図と可視化までの流れを示す. マウスログを取得するためのスクリプトを予め埋め込んだ Web ページに Web 閲覧者がアクセスする (図 2-1). アクセスするとマウスログ取得スクリプトがマウスログのフォーマットで纏めたテキストデータをログ用サーバに送信し蓄積される (図 2-2). 図 1 のフォーマットとして蓄積されたマウスログを用いて Web 閲覧者毎の特徴を掴むためにマウスカーソルの軌跡の可視化を行う. マウスログは 1 時間単位でログファイルとして蓄積されており, それらをページ毎に分割, 及び Web 閲覧者を識別しソートを行う (図 2-3).

その後, 可視化形式に合わせて座標情報を配列に変換し, JavaScript ファイルに配列として格納した座標情報を出力する (図 2-4). 出力した JavaScript ファイルを元に HTML5 の canvas 要素を用いてマウスカーソルの軌跡を点と線で描画する (図 2-5, 図 2-6). canvas 要素を, CSS を用いて iframe 要素や img 要素に重ねて表示することで, 実際のページやそのキャプチャ画像上にマウスログの可視化結果を表示する.

```
<request time>  mouse
{
    "time":<UNIX time>,"ri":<IP address>,"v":<script version>,"c" : <site code>,"h":<host>,
    "l":<request URI>,"o":<os>,"b":<brower>,"t[]":<page stay time>,"iw[]" : <innerWidth>,
    "ih[]" : <innerHeight>,"px[]":< X coordinate>,"py[]":< Y coordinate>,"id":<CookieID>
}
```

図 1: マウスログのフォーマット

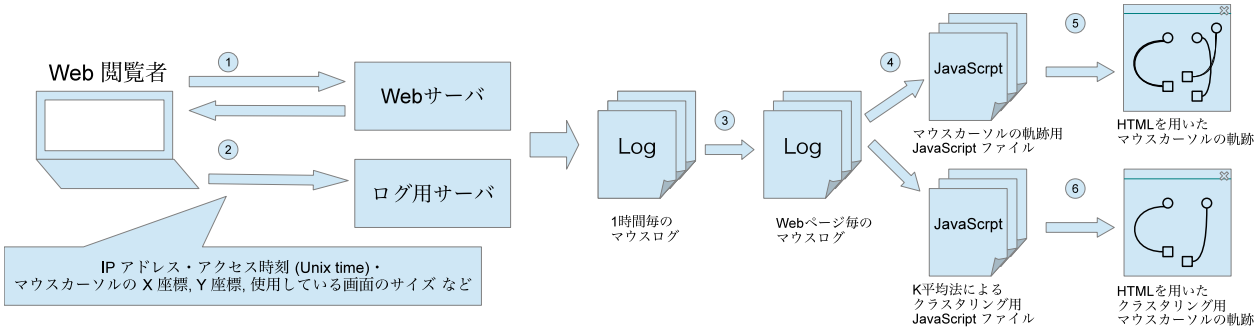

図 2: Web ビーコン型のマウスログ取得スクリプトの概要図と可視化までの流れ

3.K 平均法によるクラスタリング手法

取得したマウスログから得られる膨大なマウスカーソルの軌跡の情報から Web 閲覧者毎の特徴を掴むために, 重心線を用いた K 平均法によるクラスタリング手法を提案する. クラスタリングの手法は大きく, 階層的クラスタリングと非階層クラスタリングがあり, 非階層クラスタリングアルゴリズムである K 平均法を用いる. K 平均法は通常, 同次元のもの同士において行われるが, 本稿で扱うマウスカーソルの軌跡を構成する点要素の数はそれぞれ異なるため直接クラスタリングを行うことは出来ない. そこで K 平均法によるクラスタリングに改良を加え, 次元の異なるものに対応するために重心線を用いたクラスタリングを行う. マウスログに格納されるマウスカーソルの軌跡の本数を n, 重心線の本数を k と置き, 軌跡 $\vec{T_{i,p(i)}}$ を,

$$\vec{T_{i,p(i)}} = \vec{t_{i,1}}, \vec{t_{i,2}}, \cdots, \vec{t_{i,p(i)}} (1 \leq i \leq n, p(i) \text{ は } i \text{ により変化})$$

と定義し, 軌跡 $\vec{T_{i,p(i)}}$ を構成している $\vec{t_{i,p(i)}}$ を,

$$\vec{t_{i,p(i)}} = (x_{i,p(i)}, y_{i,p(i)})$$

と定義する. また重心線 $\vec{G_{j,m}}$ を,

$$\vec{G_{j,m}} = \vec{g_{j,1}}, \vec{g_{j,2}}, \cdots, \vec{g_{j,m}} (1 \leq j \leq k)$$

と定義し, 重心線 $\vec{G_{j,m}}$ を構成している $\vec{t_{i,p(i)}}$ を,

$$\vec{g_{j,m}} = (x_{j,m}, y_{j,m})$$

と定義する. マウスカーソルの軌跡を分類後, K 平均法を用いて集約化を行う. k 本の重心線をランダムに配置し, $p(i)$ が一定以上の全ての軌跡をもっとも近い重心線に割り当てる. 重心線を割り当てられた軌跡の平均の場所に移動し, 再び割り当てを行い, 割り当ての変更が無くなるまで最大 100 回繰り返す. 軌

跡と重心線の近さは，l=1 の場合 $t_{i,1}^{\rightarrow}$ と $g_{j,1}^{\rightarrow}$, $2 \leqq l \leqq m-1$ のとき, $t_{i,l-\frac{l}{2(l-1)}}^{\rightarrow}$ と $g_{j,l}^{\rightarrow}$, $l = m$ のとき $t_{i,p(i)}^{\rightarrow}$ と $g_{j,m}^{\rightarrow}$ の距離を足した合計値で求める．また，k, m, クラスタリングする対象とする軌跡を構成する点の数の閾値 th を任意に指定する．

4. 結果・考察

4.1. データセットを用いた場合の結果

図 3 はマウスカーソルの軌跡が 200 本 ($n = 200$) のデータセットを示す．このデータセットに対して重心線を用いた K 平均法によるクラスタリングを行った．図 4 に 200 本 ($n = 200$) のデータセットに対し重心線を 4 本 ($k = 4$) でクラスタリングした結果を示す．また図 5 に重心線を 4 本 ($k = 4$) を示す．データセットのマウスカーソルの軌跡が右上, 右下, 左上, 左下に固まっているが, 図 4 からそれぞれがクラスタ毎に分類されており, また図 5 から重心線がマウスカーソルが固まっている部分に近い座標になっているのが分かる．しかし, 左上のマウスカーソルの軌跡の重心線ではクラスタ毎に分類されたマウスカーソルの軌跡から近い座標になっていないのが分かる．これは, 重心線が 2 本あるべき場所に 1 本しかないことが原因であると考える．

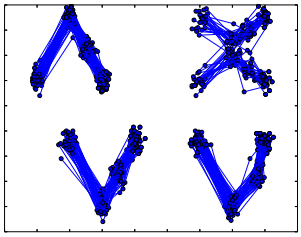

図 3: データセット ($n = 200$)

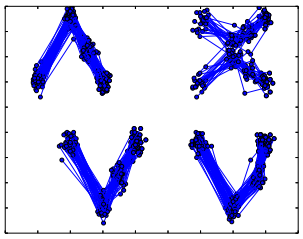

図 4: K 平均法を用いたクラスタリング ($k = 4$)

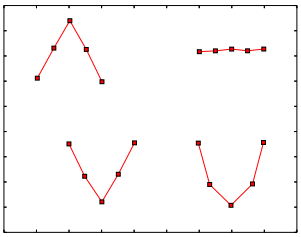

図 5: 重心線 $k = 4$

4.2. 研究室のマウスログに適用した結果

我々が運営, 管理している研究室の Web ページにログ取得スクリプトを配備し, 4ヶ月間のログデータを取得した．その集計結果を表 1 に示す．

表 1: 分析対象ログデータ

ログ取得期間	2017 年 7 月〜10 月
訪問件数 (回)	1,400
Web 訪問者 (人)	91
マウス点数 (点)	45,295

図 6 では可視化を行った研究室の Web ページを示す. この Web ページは Web サイトのトップページであることから最もアクセスが多く, 154 本のマウスカーソルの軌跡を取得した．図 7 ではマウスカーソルの軌跡のマウスログの可視化結果を示す．始点は丸, 終点は四角で示した．マウスカーソルの軌跡を時間帯毎に色付けしており, 0 時から 23 時までにアクセスした Web 閲覧者のマウスの動きを表示している．図 8 ではマウスカーソルの軌跡を $k = 3$ でクラスタリングを行った可視化結果を示す．始点は丸, 終点は四角で示しており, 集約した本数も視点と終点に示した．図 9 ではマウスカーソルの軌跡を $k = 4$ でクラスタリングを行った可視化結果を示す．始点は丸, 終点は四角で示しており, 集約した本数も視点と終点に示した．図 8, 図 9 によりこの Web ページにアクセスした Web 閲覧者は中央の写真で最もマウスカーソルを動かしていることや, ヘッダーメニューよりもメインメニューに注目してること, その中でも「 Web アプリ」「研究実績」により注目していることなどが分かる．

図 6: 可視化した Web ページ

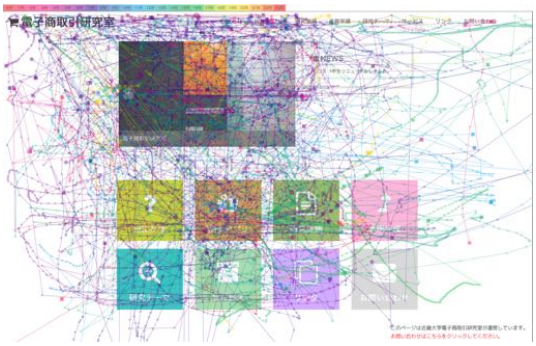

図 7: マウスカーソルの軌跡

図 8: K 平均法を用いたクラスタリング ($k = 3$)

図 9: K 平均法を用いたクラスタリング ($k = 4$)

5. おわりに

そこで本研究では, Web ビーコン型のマウスログ取得スクリプトを用いることで得られたマウスログから, アクセスしたページに対し Web 閲覧者のマウスカーソルの軌跡を可視化を行う手法に加え, Web 閲覧者がどの部分に注目するのかを知るために, K 平均法を用いたクラスタリングによる行動分析手法について提案した. 結果として K 平均法を用いたクラスタリングを行った可視化結果より, Web 閲覧者毎の注目箇所を認識可能になった. 今後の課題として, K 平均法によるクラスタリングにおいてクラスタの数を予め決めて実験を行なっている. このことによりマウスカーソルの点数が多いカテゴリでは十分に分類されるが, 一方で少ないカテゴリでは十分に分類されないという問題が挙げられる. K 平均法以外のクラスタリング手法で X 平均法という手法があり, これは最適なクラスタ数を自動で判断し分類するため, そちらを使用することで精度が向上すると考える.

参考文献

[1] Luis A. Leiva, Jeff Huang, "*Building a better mousetrap: Compressing mouse cursor activity for web analytics*, " Information Processing & Management, Vol. 51, Issue 2, pp. 114-129, 2015.

[2] Anil K.Jain, "*Data clustering: 50 years beyond K-means*, " Pattern Recognition Letters, Volume. 31, Issue 8, 1, pp. 651-666, 2010.