2F1-02

滞在時間を考慮した主要行動パターン抽出方法の検討

A study on mining of frequent activity patterns considering stay duration

服部 可奈子* 小磯 貴史* 今崎 直樹*
Kanako HATTORI Takashi KOISO Naoki IMASAKI

* (株)東芝 研究開発センター システム技術ラボラトリー
System Engineering Laboratory, Corporate R & D Center, TOSHIBA Corporation

In order to improve the profitability or the popularity, understanding customers' behavior is essential for commercial facilities. A pedestrian tracking system, which can observe trajectories of human movement, is expected to realize more precise analysis on human activities compared to the previous manpowered survey. This paper proposes a method to extract frequent activity patterns out of the trajectories considering stay duration as well as order of visiting places.

1. はじめに

PHS や GPS 機能付携帯電話の普及で市街地のような広い領域において歩行者の位置検出が可能になり，徘徊老人や営業マンの行動管理サービスが行われている[1]. その一方で, 商業施設の集客性や収益性, 顧客満足度を向上させるためには顧客行動を知ることが不可欠であり, 従来から顧客行動分析として, POS, アンケート, 出入口での来店客のカウントやビデオ映像による情報収集等が行われてきた[2]. しかし, これらの手法で得られる情報は顧客行動の限られた一側面にすぎない.

商業施設内において顧客の位置検出が可能になれば顧客の来店から退店までの移動経路を線で結び，これを他の情報と照合することにより顧客行動をより詳しく知ることができると考えられる. そこで，我々は RFID を利用した位置検出システム，歩行者動線観測システム[3]を利用して商業施設内の顧客行動を観測・分析することをめざしている. 例えば頻出する行動パターンを抽出できれば，施設の利用形態の特徴を知ることができよう.

本稿では，第一段階として行動パターンを場所と滞在時間の組からなるイベントシーケンス（以後 ES」とする）と定義し, 大量データから時間の概念を含む主要行動パターン（「主動線」と呼ぶ，以下同じ）の抽出方法を考える. 同時に, 抽出時の計算量の軽減も狙いつつ, 抽出する主動線の断続性を調節するパラメータを導入し，ユーザの目的に応じ地理制約の強弱を考慮した行動パターンの抽出手法について述べる.

2. 問題の定式化

2.1 行動空間と行動履歴データ

ある施設内で来場者が訪問できる場所（スポット）と，場所間の移動経路を，無向グラフ $G(S,E)$ で表し「行動空間」と呼ぶ. ここで，S は訪問スポット集合 $\{s_1, s_2, \cdots, s_i, \cdots, s_{n_{spot}}\}$，$E$ はリンク集合 $\{e_1, e_2, \ldots, e_d, \ldots, e_{n_{edge}}\}$ とする. 総訪問スポット数は n_{spot}，総道数は n_{edge}，各スポット間を直接接続する総経路数を e_d と表す. また，行動空間 G 内を行動する来場者を集合 $P=\{p_1, p_2, \ldots, p_j, \ldots, p_{n_{person}}\}$，総数を n_{person} と表す.

歩行者動線観測システム[3]からは，来場者毎に訪問したスポットの履歴データが取得出来る. 来場者 p_j の訪問スポット履歴データ列 a_{p_j} は次のように表される.

$$a_{pj} = \langle a_{pj1}, a_{pj2}, \ldots, a_{pjk}, \ldots a_{pjn_a(p_j)} \rangle \quad \cdots (1)$$

$$a_{p_jk} = (s(p_j,k), t_a(p_j,k), t_s(p_j,k)),\ k=1,\cdots,n_a(p_j) \quad \cdots (2)$$

連絡先：(株)東芝 研究開発センター，〒212-8582 神奈川県川崎市幸区小向東芝町 1 番地，Tel: 044-549-2408, kanako.hattori@toshiba.co.jp

ここで，$\langle\ \rangle$ は要素が生起順に並んでいることを表す. 各要素 a_{p_jk} を「訪問イベント」と呼び，訪問スポット $s(p_j,k)$ $(\in S)$，到着時刻 $t_a(p_j,k)$，滞在時間 $t_s(p_j,k)$ で構成される. 総訪問イベント数は $n_a(p_j)$ で表す.

2.2 主動線

多数の来場者に共通な行動パターン a_q（q：主動線 ID）を「主動線」と定義し，訪問イベント a_{qr} を用いて(3)のように表す.

$$a_q = \langle a_{q1}, a_{q2}, \ldots, a_{qr}, \ldots, a_{qn_a(q)} \rangle \quad \cdots (3)$$

r は主動線として抽出された訪問イベントの生起順番号，$n_a(q)$ は a_q の要素数である. a_{qr} は(4)で定義する.

$$a_{qr} = (s(q,r), \tilde{t}_s(q,r)) \quad \cdots (4)$$

ただし $\tilde{t}_s(q,r)$ は訪問スポット $s(q,r)$ における滞在時間を表すファジィ変数で，メンバーシップ関数で予め定義される. (3), (4)により, 主動線は到着時刻を考慮せず, かつあいまいな滞在時間をもつ訪問イベント（これを「行動イベント」と呼ぶことにする）の訪問順シーケンスとして表現される. 総主動線集合は $A=\{a_1, a_2, \ldots, a_q, \ldots a_{n_{fmp}}\}$ とする. n_{fmp} は総主動線数を表す.

主動線を構成する各行動イベント間には，行動イベントが起こるスポット間の地理的制約と，その生起に対する高い相関が存在すると考えられる. そこで, 主動線を以下の二つの条件を満たすものとして定義する.

【条件 1 地理制約】

主動線中の隣り合う 2 行動イベントの間に許容できる訪問スポット数を c (≥ 0) とする. この c を地理制約パラメータと呼ぶ.

行動空間 $G(S,E)$ において，2 訪問スポット間を直接つなぐ道数を表す行列 M_0 は以下の n_{spot} 次元正方行列で表される. $m_{s_f s_g}$ は s_f, s_g を両端点とする道の本数である.

$$M_0 = \begin{bmatrix} m_{s_1s_1} & \cdots & m_{s_js_1} & \cdots & m_{s_{n_{spot}}s_1} \\ \vdots & \ddots & & & \vdots \\ m_{s_1s_j} & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ m_{s_1s_{n_{spot}}} & \cdots & \cdots & \cdots & m_{s_{n_{spot}}s_{n_{spot}}} \end{bmatrix} \quad \cdots (5)$$

いま，ノード数 c 以下を経由するリンク数を行列 M_c で表すと次式で書ける.

$$M_c = \sum_{l=0}^{c} M_0^{l+1} \quad \cdots (6)$$

このとき，主動線 $a_q = \langle a_{q1}, a_{q2}, \cdots, a_{qr}, \cdots a_{qn_a(q)} \rangle$ に含まれるすべての $s(q,r)$ は次式を満たさなくてはならない.

$$\begin{aligned} &m^c_{s(q,r)s(q,r+1)} \neq 0 && \text{if } r=1 \\ &m^c_{s(q,r)s(q,r+1)} \neq 0 \wedge m^c_{r(r-1)} \neq 0 && \text{if } 1<r<n_a(q) \\ &m^c_{s(q,r)s(q,r-1)} \neq 0 && \text{if } r=n_a(q) \end{aligned} \quad \cdots (7)$$

【条件2:滞在時間を考慮した出現率】

訪問スポットごとの滞在時間の滞在時間集合を M_{si} で定義する. 例えば行動履歴データから訪問スポットごとの滞在時間分布を作成し, 目視で分割点を指定し, 分割点を基に $m_{s_f}(t_s)$ を作成する. 作成するメンバーシップ関数の総数を n_{msfi} とすると $M_{s_i}=\{m_{s_1}(t_s),\cdots,m_{s_f}(t_s),\cdots,m_{s_{n_{msfi}}}(t_s)\}$ となる.

すると, 主動線候補 a_q が全来場者の行動に現れる出現率は,

$$\mathrm{support}(a_q)=\frac{\sum_{j=1}^{n_{person}} f_q(a_{pj},a_q)}{n_{person}} \quad \cdots(8)$$

で示される. ただし, a_{pj} に含まれる a_q が(7)を満たすとき $f_q(a_{p_j},a_q)=\min_s m_{s_f}(t_s(p_j,k(s_i)))$ とし, それ以外のときは $f_q(a_{pj},a_q)=0$ とする. ここで, $k(s_i)$ は訪問イベント a_{p_jk} において $s_i=s(p_j,k)$ となる k を表す.

このとき, 主動線 a_q は出現率に関し設定される閾値 min_support (∈[0,1])に対して, (9)を満たさなくてはならない.

$$\mathrm{support}(a_q)\geq \mathrm{min\_support} \quad \cdots(9)$$

3. 主動線の抽出

Agrawal の相関抽出法[4]をベースに, 前節で定義した主動線の抽出を行うアルゴリズムを以下に示す.

Step1: min_support, c をそれぞれ設定する.

Step2: 訪問スポットごとに滞在時間集合 M_{si} を定義する.

Step3: 主動線候補集合 $A_1=\{a_{x(1)}\}$ を検索する. ただし, $a_{x(1)}$ は $n_a(q)=1$ において $a_{x(1)}=(s_i,\tilde{t}_s(\ell))$ となるすべてのイベントである. $\tilde{t}_s(\ell)$ は Step2 において分割された m 番目のファジィ滞在時間を表す. A_1 の中で, (9)を満たすイベントシーケンス (ES) を抽出し, これを主動線集合 A'_1 とする.

Step4: $n_a(q)=2$ として $A_2=\{a_{x(2)}\}$ を検索する. ただし, $a_{x(2)}=\langle a'_{x1},a'_{x2}\rangle$ は $a'_{x1}\in A'_1\cap a'_{x2}\in A'_1$ であり, かつ a'_{x1},a'_{x2} が(9)を満たすすべての ES である. A_2 の中で式(9)を満たす ES を抽出し, 主動線集合 A'_2 とする.

Step5: $n_a(q)=n_a(q)+1$ として $A_{n_a(q)}=\{a_{x(n_a(q))}\}$ を検索する. $a_{x(n_a(q))}$ は, $a_{x(n_a(q)-1)}\in A'_{n_a(q)-1}$, $a_{y(n_a(q)-1)}\in A'_{n_a(q)-1}$ の部分集合, $\langle a'_{x1},a'_{x2},\ldots,a'_{xna(q)-2}\rangle$ と $\langle a'_{y2},a'_{y3},\ldots,a'_{yna(q)-1}\rangle$ が等しいとき, $a_{x(n_a(q))}=\langle a'_{y1},a'_{x1},a'_{x2},\ldots,a'_{xna(q)-1}\rangle$ をすべての ES に対して作成する. $A_{n_a(q)}$ の中で(9)を満たす ES を抽出し, 主動線集合 $A'_{n_a(q)}$ とする.

Step6: $A_{n_a(q)}\neq\varnothing\cap A'_{n_a(q)}\neq\varnothing$ の場合 **Step5** に戻る.

得られる主動線集合は $A=\{A'_1,A'_2,\ldots,A'_{n_a(q)-1}\}$ である.

4. 実験

提案手法の検証を行うため, 図1のような A～P の16箇所の観光地を廻る歩行者 1000 人を想定して, 擬似的に行動履歴データを生成する. 滞在時間は各訪問スポットについて短・中・長に3分割する.

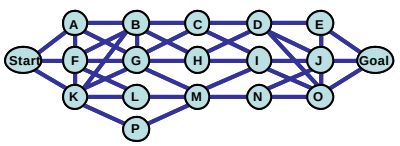

図 1. 想定した観光地

提案手法による主動線抽出結果を, 相関抽出法の結果とともに表1に示す. 設定は min_support=0.2, c =0, 1, 2,∞(地理制約を考慮しない)とした. 同表は, 抽出された $n_a(q)$ =3 の主動線のうち, support が高いものを順に3つ示している.

表 1. 抽出された主動線

	抽出された主動線数(support数順序)		
	1	2	3
相関抽出法	(G, H, I)	(H, I, J)	(G, H, J) または (G, I, J)
c=0	(B(長), C(短), D(短))	なし	なし
c=1	(F(短), K(短), M(短))	(B(長), C(短), D(短))	(F(短), M(短), O(長))
c=2	(F(短), K(短), M(短))	(B(長), C(短), D(短))	(F(短), M(短), O(長))
c=∞	(F(短), K(短), M(短))	(F(短), K(短), O(長))	(B(長), C(短), D(短))

相関抽出法と本提案手法との違いを考察する. まず抽出された support の上位3つの主動線がまったく異なる. これは G,H,I,J の訪問数は多いもののその滞在時間分布が一様だったのに対し, (B,C,D)や(F,K,M)では滞在時間分布に一定の偏り傾向があったからである. 提案手法では, (B(長),C(短),D(短))や(F(短),K(短), M(短))等を抽出できた.

また, 同じ min_support 値を与えたとき, 提案手法では相関抽出法と比べ抽出される主動線数が少ない. これは提案手法では滞在時間を考慮するため support 値が一般に小さくなってしまうからである. そのため, 提案手法を用いる場合は min_support 値を相関抽出法よりも小さく設定する必要がある.

次に地理制約パラメータ c について考察する. 一般に, c の値を増加させると得られる主動線の種類が増加する. c =0 のときに抽出される主動線は(B(長), C(短), D(短))のみである. c =0 は, スポット間が直接道で接続されていなければならない非常に厳しい制約である. 一方, c =∞のとき抽出された主動線(F(短) ,K(短), M(長))は K と M の間が地理的に離れており, その間の経路が非常に多く考えられる. このような主動線を主動線と捉えるのかはユーザにより異なる. c の値を小さくすると経路が重視され, c の値を大きくするとスポット間の相関性が重視される. これらのことより, ユーザが c を調節することで, 抽出する主動線の性質を調節できると考えられる.

なお, c の値を小さくすると $A_{n_a(q)}$ の数を絞り込む際の計算量が軽減されることを確認した. 例えば, c =0 のとき主動線候補集合 A_2 の総数が, c =∞のときの3分の1となった.

5. 結論

提案手法で行動履歴データから主動線を抽出できることが実験により確認できた. 本手法は滞在時間を考慮するとともに, 主動線の性質を地理制約パラメータで調節することができるため, 実用的な行動マイニング手法の足がかりになると考えられる.

今後, 有効性を検証するため実データでの実験を早急に行う予定である.

参考文献

[1 navi-p 2003] (例えば)ナビッピドットコム株式会社:位置情報 ASP サービス, http://www.navi-p.com/gpsasp/index.html

[2 Envirosell 2002] (例えば)エンバイロセルジャパン株式会社, http://www.envirosellj apan.com/index.html

[3 Imasaki 2001] 今崎 直樹: 歩行者動線観測システムと展示会会場混雑予測モデルの構築, 第 11 回 FAN シンポジウム予稿集, pp.509-514, 2001.

[4 Agrawal 1994] R.Agrawal: Fast algorithms for mining association rules in large databases, In Proc. Of the 20th Int'l Conf. on Very Large Data Bases (VLDB) , pp.478-499, June 1994.