Web 利用パターンマイニングのための
アクセスシーケンスの導出

知能系　和田・犬塚研究室
No. 16115145　向井 啓樹

1　はじめに

Web アクセスログに注目したデータマイニングは Web 利用マイニングと呼ばれ，Web ユーザの振る舞いを理解することを目的とする．早川ら [1] は，各ページへのユーザのアクセスの道のりを**アクセスシーケンス**と呼び，Web アクセスログの解析によってアクセスシーケンスのマイニングを系列データマイニングに帰着できることを示した．本研究では，早川ら [1] の手法を実際の Web アクセスログに適用した場合の問題点発見とその改善手法の提案を目的とする．

2　アクセスシーケンス

Web アクセスログにおいてユーザ x のログの ID の列を R_x とし，これを節の集合とし，辺の集合を次の E_x とする有向グラフ $G_x = (R_x, E_x)$ を考える．

$$E_x = \{(i,j) \in R_x^2 \mid (request(i) = referer(j)) \wedge (request(i) \neq referer(k)), i < k < j\}$$

$referer$(n) は n 番目のログの参照元 URL，$request$(n) はリクエスト URL を表す．早川ら [1] は，G_x を**ログレコードグラフ**と呼び，ログレコードグラフにおいて入次数 0 のノードから出次数 0 のノードまでの有向路をアクセスシーケンスと定義した．

3　アクセスシーケンス導出アルゴリズム

本研究では，一度だけのデータスキャンでアクセスシーケンスを導出するアルゴリズムを与えた (図 1)．

Input: ユーザ x の Web アクセスログのリスト R_x(日時が新しい順にソート済)
Output: ユーザ x のアクセスシーケンスのリスト $SequenceList$

```
1.  SequenceList = [],Open = []
2.  foreach r1 ∈ R_x
3.      Open に referer(r1) を追加
4.      if request(r1) ∈ Open
5.          Open から request(r1) を削除
6.          foreach Sequence ∈ SequenceList
7.              r2 ← SequenceList の先頭要素
8.              if referer(r2) = request(r1)
9.                  Sequence の先頭に r1 を追加
10.     else
11.         SequenceList に [r1] を追加
12. return SequenceList
```

図 1: アクセスシーケンス導出アルゴリズムの疑似コード

4　間接リクエストの除去

本研究では，実際の Web アクセスログには，ユーザの振る舞いに起因するログと，ページ内の画像読み込みといった自動アクセスに起因するログが混在している点に注目し，前者を**直接リクエスト**，後者を**間接リクエスト**と呼ぶことにする．間接リクエストはユーザの意図的な振る舞いでないため，図 1 で与えたアルゴリズムは間接リクエストを除去していない点が問題である．そこで，ログのリクエスト日時差に着目することで間接リクエストを除去したアクセスシーケンスを導出するアルゴリズムを提案する．新たに閾値 $SeeTime$ を設け，次の R'_x を節の集合とする有向グラフ G'_x をログレコードグラフとして再定義する．

$$R'_x = \{i \in R_x \mid G_x\text{で節 } i \text{ から出次数 0 の節 } j \text{ への有向路に閲覧辺 } (k,l) \text{ がある}, i \leq k < l \leq j\}$$

閲覧辺とは，リクエスト日時差が SeeTime 以上である節のペアを結ぶ有向辺を示す．本研究では，この G'_x からアクセスシーケンスを導出する改善アルゴリズムを示した．この改善アルゴリズムでは，間接リクエストを除去することによりユーザの振る舞いでないアクセスシーケンスが除去されている．この改善アルゴリズムも必要なデータスキャンは一回だけである．

5　実験

3 節と 4 節それぞれで示されたアルゴリズムによって導出されるアクセスシーケンス数を比較した (図 2)．実験には，プロキシサーバに蓄積された本研究室の学生らの Web アクセスログを用いている．

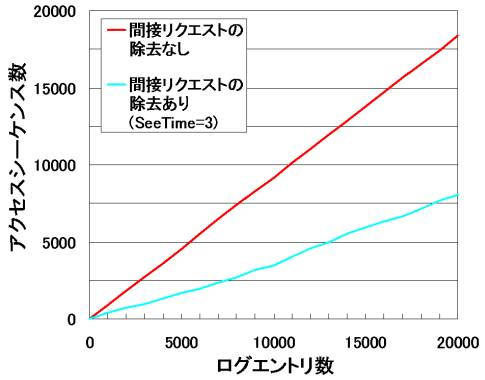

図 2: 導さ出れるアクセスシーケンス数の比較

6　まとめ・今後の課題

本研究では，一回のデータスキャンでアクセスシーケンスを導出するアルゴリズムを与え，さらに間接リクエストの除去を考慮した改善アルゴリズムを提案した．また，IP 番号によるユーザ分類の妥当性を示した．今後の課題として，リクエスト日時差の考慮によって本当に間接リクエストが除去されているかといった検証が挙げられる．

参考文献

[1] N.Inuzuka and J.Hayakawa. A Unified Approach to Web Usage Mining Based on Frequent Sequence Mining: KES 2007, Springer, pp987-994, 2007.