

学習履歴の時分割による多重仮名化に関する研究

1200299 大林久也 【教育情報工学研究室】

1 はじめに

個人情報とは、特定の個人を識別可能な情報のことであり、それを保護する技術として匿名化と仮名化がある。LA (Learning Analytics) の研究のために LRS (Learning Record Store) に蓄積される学習履歴には個人を特定可能な情報が含まれており、個人情報を保護する技術を施さなければならない。しかし、匿名化を行うと詳細な分析が困難になってしまうため仮名化の方が望ましい。仮名化とはデータセット内のあるユーザの識別子を、別の文字列に置き換える技術である。また、1つの識別子に対して複数の仮名に置き換える仮名化を多重仮名化という。京都大学の LA 環境では、学習履歴内のユーザの識別子を LMS の UUID (Universal Unique Identifier) に置換してから LRS に蓄積している [1] が、仮名の更新は行われていない。仮名を更新せず学習履歴を蓄積し続けていると、個人の特徴や傾向を掴むことが容易になり個人特定の可能性が高まる。そこで本研究では、LRS に蓄積される学習履歴の時系列による多重仮名化を行い、その定量的評価を行う。

2 仮名更新頻度のシミュレーション

本研究では、データの匿名加工と再識別の性能を競うコンテストである PWS CUP 2016 で多くのチームが用いた Jaccard 係数による再識別手法を参考として仮名更新頻度の評価を行う。同一ユーザに割り当てられた仮名が平均してどの程度推測されるかを評価するために、更新頻度ごとの Jaccard 係数の平均値を類似度の評価指標として用いる。Jaccard 係数は、2つの異なる仮名 p, q 、各仮名に対応する URI の集合を $Access(p)$ 、 $Access(q)$ とすると (1) で求められる。

$$Jaccard(p, q) = \frac{|Access(p) \cap Access(q)|}{|Access(p) \cup Access(q)|} \quad (1)$$

2.1 使用データ

68人のユーザから収集された、2019年12月5日から12月26日までの情報科学3の学習履歴を用いる。総数は30328個であり、1人当たりの平均は411.29個、1人当たりの最大は2360個である。学習履歴は3つのステートメント actor, object, timestamp からなり、それぞれ人物、学習物(資料のURI)、履歴が記録された時間を表す。

2.1.1 時分割による仮名更新

仮名の更新頻度を1週間に1回または2回、時間割に沿った時間帯、24, 12, 8, 6, 4, 3, 2, 1, 0.5 時間に設定した。週1回更新では日曜日または月曜日の2通り、週2回更新では日曜日と水曜日、月曜日と木曜日の2通

りに更新を行ったため合計14通りである。時間割に沿った時間帯とは00:00~8:44, 8:45~10:24, 10:25~12:04, 12:05~12:54, 12:55~14:34, 14:35~16:14, 16:15~17:54, 17:55~23:59という時間帯である。どの更新頻度でも00:00を基準とする。

2.1.2 学習履歴ごとの仮名更新

まず、ユーザ1人あたりに割り振る仮名の個数をあらかじめ1, 2, 3, 4, 5, 7, 9, 12個に設定しておき、その個数分だけ各ユーザの仮名を生成する。そして、学習履歴が記録されるごとに、生成した仮名を順番またはランダムに割り振っていき仮名を更新していく。

3 結果と考察

表1, 2はそれぞれの更新頻度における Jaccard 係数の平均値と総仮名数の一部を示したものである。「全通り」とは異なる2つの仮名の組全通り、「同一ユーザ」とは同一ユーザに対応する仮名の組全通りの Jaccard 係数の平均値を表す。Web アクセス履歴の仮名更新頻度に関する研究 [2] と比べて値が高いが、これは1つの講義だけの学習履歴を用いたため、各仮名に対応する URI がほぼ同じであることが原因であると考えられる。

表1 時分割による仮名更新の Jaccard 係数の平均値

更新頻度	全通り	同一ユーザ	総仮名数
週1(日曜日)	0.46191	0.32519	164
週2(日・水曜日)	0.46024	0.35424	209
24時間	0.45735	0.45735	218
時間割	0.46230	0.42237	289
0.5時間	0.48039	0.47741	854

表2 学習履歴ごとの仮名更新の Jaccard 係数の平均値

生成した仮名数	全通り	同一ユーザ	総仮名数
2個(順番)	0.63758	1.00000	136
2個(ランダム)	0.63847	0.99705	136
12個(順番)	0.64856	0.97214	810
12個(ランダム)	0.65329	0.95951	807

4 まとめ

本研究では、LRS に蓄積される学習履歴の仮名化更新頻度についてシミュレーションを行った。今後は複数科目を対象とした多重仮名化の評価を行いたい。

参考文献

- [1] Brendan Flanagan, Hiroaki Ogata, "Integration of Learning Analytics Research and Production Systems While Protecting Privacy", Proceedings of the 25th ICCE, 2017.
- [2] 福嶋雄也, 北島祥伍, 満保雅浩, "仮名の再同定率に基づく仮名更新頻度の定量的評価", 情報処理学会論文誌, Vol.60 No.9 1509-1517 2019.