匿名化ログ分析における複数属性の等価性再識別リスクと分析効率とのトレードオフ評価

2017SC075 髙見将宗
指導教員：石原靖哲

1 はじめに

近年，ログ所有者がデータベースサービスに暗号化ログデータベースの管理を委ねるといった形態が増えつつある．この時，ログデータは厳重に扱う必要があるため，ログ所有者によって暗号化されている．ログ管理は，主にログ収集，保管，分析などから構成される．分析の際は，暗号化ログデータベースを分析するツールを用いて，追跡したいイベント情報を突き止める．そして，ログ所有者が提供しているサービスにその情報を迅速に反映させる．この時，ログ管理者は，個人特定のリスクを避けるために不用意に個人情報を知る必要はない．また，サービスに情報を迅速に反映させる面から効率的分析も要求される．

上述のような匿名化ログ分析の問題を，ログ所有者とログ管理者で行う．ログ所有者は，秘匿対象ログを匿名化し，そのデータベースをログ管理者に委託する．ログ所有者は，一般にログ管理者より計算能力が弱いため，ログの分析をする際にはログ管理者に対してログデータの等価性情報の一部を開示しつつ，分析作業の一部を委託する．そしてログ管理者から受け取った結果を利用して，分析作業を完遂する．この時，ログ管理者への等価性情報の開示量が多いほど全体として分析効率が上がるが，ログ管理者に等価性を再識別されるリスクが上がる．

先行研究では，ユーザ ID 等価性を用いた匿名化ログ分析を提案している [3]．本研究では，複数属性等価性を用いた匿名化ログ分析の手法を提案する．また，2 パターンのタグ生成法を提案し，それらのタグについて，等価性再識別リスクと分析効率とのトレードオフ評価をする．

2 設定

2.1 対象ログデータ

本研究で使用したデータセットには，ウェブオンラインショッピングにアクセスしているユーザのアクセスログが収集されている [2]．このアクセスログから 1 万ログを対象に分析を行う．また，本研究では IP アドレスとユーザエージェントを匿名化対象の属性とする．

2.2 データベース分析ツール

本研究では，ログデータベース分析ツールとして InfluxDB を採用する [1]．InfluxDB はオープンソースの時系列データベースシステムの一つで，時系列データを格納するのに適しており，アクセスログの集計や解析が可能となっている．

2.3 検知対象イベント

本研究では，IP アドレス + ユーザエージェント毎の単位時間当たりのアクセス回数が閾値を超えるログデータを対象に検知する．また，ウェブオンラインショッピングにアクセスしているユーザのアクセス回数を確認し，閾値以上のユーザをヘビーユーザとして検知することを本研究でのシナリオとする．

3 トレードオフ評価のための方式

3.1 等価性情報を一部開示するタグ

3.1.1 集約化タグ

A と B を個人情報の平文とする．それぞれから生成したタグを T_A，T_B とする．集約化タグとは，$T_A \neq T_B$ ならば $A \neq B$ が成立するタグである．本来 z 種類存在する個人情報に対し，x 種類 $(x < z)$ のタグを確定的に生成することで，集約化タグを生成する．

3.1.2 細分化タグ

A と B を個人情報の平文とする．それぞれから生成したタグを T_A，T_B とする．細分化タグとは，$T_A = T_B$ ならば $A = B$ が成立するタグである．あらかじめ y 種類 $(y > 1)$ の鍵を用意し，ログレコード毎に擬似ランダムに鍵を選択しハッシュ化することで，y 種類の細分化タグを生成する．

3.2 2 種類のタグを用いたログ分析

2 種類のタグを用いて，以下の 2 ステップでログ分析を行う．

1 ステップ目では，集約化タグをキーとし，検索クエリを実行する．集約化タグ毎の単位時間当たりのアクセス数を DB 分析ツールから取得する．本研究では，アクセス数が閾値 20 回を超えた集約化タグを取得する．問合せでは，異なる IP アドレス + ユーザエージェントが同じタグとなっているため，この段階でアクセス数の閾値を超えていないグループは細分化タグでの検知対象から除外することができる．

2 ステップ目では，前ステップでの問合せ結果のログデータに対して，細分化タグをキーにし，検索クエリを実行する．細分化タグ毎の単位時間当たりのアクセス数を DB 分析ツールから取得する．

4 複数属性のための分析手法

本節では，2 パターンのタグ生成方法に基づいた複数属性のための分析手法を提案する．

表 1　2 種類のリスク R_x，R_y

	複数属性のタグ	属性毎のタグ
R_x	$1-\frac{1}{x}$	$1-\frac{1}{x_1}\times\frac{1}{x_2}$
R_y	$\frac{1}{y}$	$\frac{1}{y_1}\times\frac{1}{y_2}$

4.1　複数属性のタグを生成する方法

IP アドレスとユーザエージェントの平文を繋げ，集約化タグと細分化タグを生成し付与する．利点は，タグのデータ領域が少ないことである．また欠点は，IP アドレスだけの場合とユーザエージェントだけの場合の分析はできないことが挙げられる．

4.2　属性毎のタグを利用する方法

IP アドレスとユーザエージェントからそれぞれ集約化タグと細分化タグを生成し付与する．ペアとなる IP アドレスのタグとユーザエージェントのタグから分析を行う．利点は，複数属性だけでなく IP アドレスだけの場合とユーザエージェントだけの場合の分析も可能となる．欠点は，タグを保存するためのデータ領域が多いことが挙げられる．

4.3　複数属性等価性再識別リスク

複数属性等価性が再識別されるリスクの式を表 1 に示す．x，x_1，x_2 は集約化パラメータ，y，y_1，y_2 は細分化パラメータである．R_x は，集約化タグの開示により 2 つの平文が異なることを識別されるリスクである．R_y は，細分化タグの開示により 2 つの平文が同じであることを識別されるリスクである．

5　実験結果・評価

属性毎のタグ生成においては，$x=20$，30，50，70 と各集約化パラメータにおいて $y=5$，10，15 を試した．複数属性のタグにおいては，属性毎のタグのパラメータをそれぞれ 2 乗して試した．結果を図 1，2 に示す．

図 1 を見ると，複数属性のタグと属性毎のタグそれぞれにおける集約化タグでの検索時間が増加している．リスクをとることで分析効率が上がることが期待されたが，結果はリスクをとるにも関わらず集約化タグでの分析効率が下がった．これは，問合せに含まれる GROUP BY での計算に時間を要したためであると考えられる．また，図 1，2 を見ると基本的に属性毎のタグでの検索時間が複数属性のタグの検索時間を上回っている．しかし，図 1 では，複数属性のタグにおける集約化タグでの検索時間が属性毎のタグにおける集約化タグでの検索時間を $R_x<0.999$ の時に下回っているのが分かる．そして，図 2 では，R_y が増えるほど複数属性のタグでの検索時間と属性毎のタグでの検索時間の差が小さくなっている．このことから，R_x を減らし R_y を増やす方向でパラメータを調整することで，属性毎のタグと複数属性のタグの検索時間の差を減らせる可能性があると分かる．そうすることで，利便性と効率性の観点から，属性毎のタグを用いるのが良いパラメータがあると予想される．

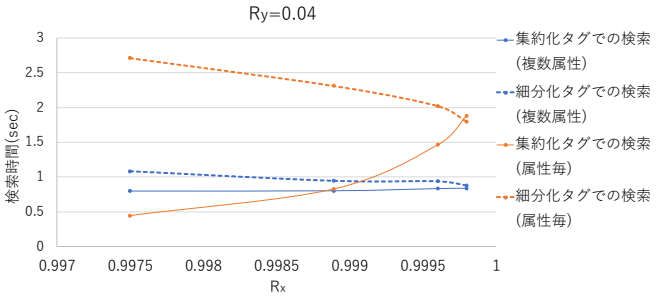

図 1　R_x と検索時間におけるトレードオフ評価

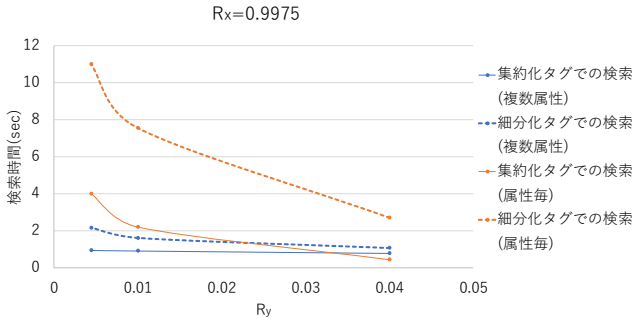

図 2　R_y と検索時間におけるトレードオフ評価

6　まとめ

本研究では，匿名化ログ分析における複数属性等価性再識別リスクと分析効率とのトレードオフ評価を行った．実験に用いたパラメータの範囲においては，複数属性のタグと属性ごとのタグそれぞれにおける集約化タグでの検索でリスクが増えるにも関わらず時間を要した．また，属性毎のタグは，複数属性のタグに比べて検索時間を要していたが，R_x を減らし R_y を増やす方向でパラメータを調整することで，2 つのタグの検索時間の差を減らせる可能性があると分かった．これからの課題として，暗号化と復号の実装をすることが必要である．また，4.3 節の属性毎のタグにおける式は，場合分けができると考え改善する必要がある．

参考文献

[1] InfluxDB. InfluxData Inc, 2020. `https://portal.influxdata.com/downloads/`.

[2] Zaker and Farzin. Online Shopping Store - Web Server Logs. Harvard Dataverse, 2019. `https://doi:10.7910/DVN/3QBYB5`.

[3] 萩尾玲太. 匿名化ログ分析におけるユーザ ID 等価性の再識別リスクと分析効率とのトレードオフ評価. 大阪大学大学院情報科学研究科修士学位論文, 2018.