2009-13-03

サービス指向型ルータに利用可能な低コストプライバシ保護手法

Low-cost Privacy Preserving Method for Service-oriented Router

80816878　橋岡大地(Daichi Hashioka)　Supervisor：西宏章(Hiroaki Nishi)

1　緒論

ルータはIPネットワークの中心に位置し，主にパケット転送とプロトコル変換を担う．一方で，高度な計算処理能力を有しながら，アプリケーション層に関与することはない．ルータが取得できるデータはエンドホストでは得難いユニークな情報を含んでいると考えられる．その理由は次の通りである．エンドホストは，例えば検索エンジンのクローラといった能動的な情報取得のみに頼らざるを得ない．それに対しルータは膨大な量のパケットが通過するため，受動的な情報収集が可能であり，情報取得のリアルタイム性にも優れている．ルータが取得するユニークな情報を外部に提供することで，ウェブサービスをよりリッチにする可能性があるといえる．

本研究の目標は，実際のトラフィックデータに基づく情報を提供するサービス指向型ルータ(Service-oriented Router: SoR) の実現に伴い顕在化すると考えられる，ネットワークトラフィック中に含まれるプライバシ情報を保護する手法の提供である．

2　サービス指向型ルータ SoR

我々の提案するサービス指向型ルータ（Service-oriented Router: SoR）[1] は，パケットストリームを監視し，ユーザの指定した抽出条件に沿ったデータをルータ内部のデータベースへ格納するルータである．図 1 に SOR における処理の概要を示す．SoR は単にデータ転送とプロトコル変換を行うのではなく，トラフィック中のパケットペイロードを監視し，Web サービス，クライアント，周辺ルータが利用可能な形で情報を提供する．トラフィックデータから抽出するデータの指定には，柔軟な正規表現(regular expression)を使用可能である．

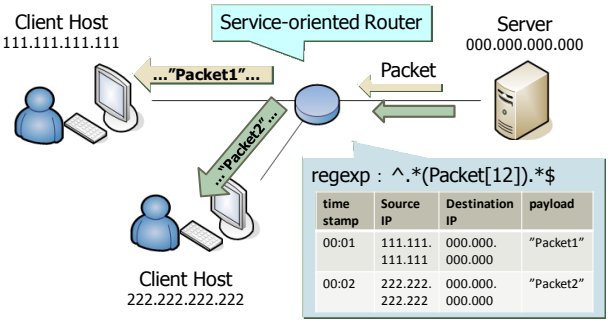

図 1　サービス指向型ルータの概念図

3　SoR におけるプライバシ情報

トラフィックデータにはプライバシ情報が含まれるため，SoR がトラフィックの中身を公開する際には，適切なプライバシ保護機構によりある程度匿名化されていることが不可欠である．センサネットワークや様々な個人情報を扱うデータベース，行動履歴解析などのアプリケーションにおいても，プライバシの問題が重要視されており頻繁に議論されている．SoR ではこの点を考慮し，個人を直接特定できる可能性のあるIPアドレス等の情報は，インターネットで公開されているバックボーンネットワークトレースデータと同様，ハッシュ化により保護する．さらにデータのペイロード部分とそれ以外の部分との関係も考慮し，データ操作によってプライバシを保護する．データ操作については，あらかじめ形式の決まったデータについては，例えば図2に示すような一般化階層で表わされる一般化関数を定義して適用する．一方，データ解析者が要求する抽出条件やネットワークトラフィック上に現れるデータを予測するのは困難である．そのため，ユーザの条件指定により抽出されたデータに限っては，加工を行わない．

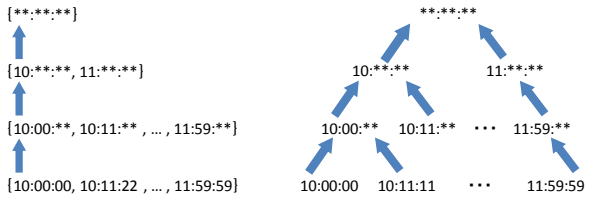

図 2　一般化階層の例

また本提案では，SoRによって得られるデータを用いたサービス/アプリケーションが提供する情報に可能な限り柔軟性を持たせるため，SoRが備えるプライバシ保護機構は様々な要求に答えつつ，一般ユーザでも設定が容易であることが望まれる．一方で，SoRを管理する側には汎用性の高いプライバシ制御設定を提供することも必要である．また，過度なプライバシ保護は提供可能な情報の価値を低減させるため，データ解析者には最低限のプライバシを維持しつつデータマイニングが可能である環境を提供することが求められる．さらに，SoRは情報提供のリアルタイム性も重要視するため，プライバシ保護処理に必要なコストが少ないのが望ましい．

このような点を考慮し，本研究では各属性の値を一般化して l-多様性を保証することによって，データテーブルのプライバシ保護を達成する．l-多様性とは，任意の整数 l について，「データテーブル中の任意の属性値を有するタプル組合せにおいて，出現する注目属性の値が少なくとも l 個の多様性を有する性質」と定義される．具体的には、データテーブル中の任意の属性値を有するタプル組合せにおいて注目属性の値が n 種類出現する場合に，以下の式が満足されるならば，そのデータテーブルは l-多様性を満足する．

$$r_1 \le c(r_l + r_{l+1} + \ldots + r_n), \quad 1 < l < n \qquad (1)$$

ここで，r_1 は最も高い出現確率を有する注目属性の確率であり，r_i は i 番目に高い出現確率である．定数 c の値は任意に設定できるが，本研究での評価においては $c = 1$ とする．データテーブルが l-多様性を満足することにより，データテーブル中のタプルと個人は一意に識別できない．この性質により，プライバシ保護が達成されていると定義する．

4　連関規則を用いたプライバシ保護手法

「属性(群) A を持つ観測対象は属性(群) B も持つ傾向にある」という知識を「連関規則」呼ぶ. A はルールヘッド(rule head)， B はルールボディ(rule body) である．A とB を確率事象とみなし，多量のトランザクション(transaction:取引)データの中から有用な連関規則を抽出する処理をバスケット分析と呼ぶ.

連関規則を用いたプライバシ保護手法では，データテーブルが l -多様性を保証するために，まずバスケット分析を行い，匿名化すべきタプルを探索する．このときトランザクションとなるのはデータテーブル中に存在する全ての値の中で，任意の属性の組合せに属する値の集合である．例えば，ある属性値 v と注目属性の値 s について，条件付き出現確率 $p(s \mid v)$ が $1/l$ を超えるタプルが存在する場合にはデータテーブルの l -多様性が保証できないため，当該タプルについて一般化処理を行う．これは，一般化階層において，一般化処理を施すほど値が最大一般化値に向かって集約されていくため，一般化の過程で別の値を有するタプルの値と集約される可能性があるためである．例えば，タプルt1 における非注目属性値と注目属性値が $v1$ と $s1$，タプルt2 においては $v2$ と $s2$ であり，データテーブルについて条件付き確率 $p(s1 \mid v1) = p(s2 \mid v2) = 1$，かつ，一般化関数 f について $f(v1) = f(v2) = v'$ であるとする．このとき，これらのタプル組について一般化を行った場合，そのデータテーブルについて $p(s1 \mid v') = p(s2 \mid v') = 1/2$ となり，それらタプルについては 2 -多様性が保証された状態となる．このように，連関規則を抽出し，特定されたタプルに対して一般化を行う操作を繰り返せば，最終的にデータテーブル全体の l -多様性が保証できる.

5　非階層的クラスタリングによる低コスト化

連関規則のみを用いた手法の場合，考慮すべき属性の数が増えるほど探索すべき連関規則の組合数も増大し，タプルの選択に必要な計算量が発散する．そこで，データテーブル中のタプルについて，非注目属性と注目属性の値を用いて非階層的クラスタリングを行う．この際，クラスタリング結果を参考にして，類似度の高いタプル同士についてあらかじめ一般化を行っておく．ある程度一般化を行った後のデータテーブルに対して連関規則を用いた手法を適用することによって，最終的に l -多様性を満たすデータテーブルを得る.

クラスタリングとは，互いに類似する観測対象同士をグループ(クラスタ)化する手法であり，本提案では非階層的クラスタリングのアルゴリズムとして代表的な K -means法を用いる．クラスタリングを行った結果得られたクラスタに対して一般化処理を行うのは，あるタプル組の非注目属性値について一般化を行った結果，そのタプル組が l -多様性を満たす可能性があるためである.

クラスタリング処理は考慮する属性の数によって計算量が増加しない特長がある．また，クラスタリング処理を行った後であれば，連関規則で探索すべきルール数が減少する．結果，データ量や考慮すべき属性が増えた場合でもより少ない計算時間内での処理が期待できる.

6　評価

それぞれの手法において, タプル数(データ数)の増加による実行時間の傾向と l 値の変化による実行時間の傾向を評価した．図 3 と図 4 はプライバシを考慮する属性数が 2 の場合の評価結果である．連関規則を用いた手法は，l 値や考慮すべき属性数が小さい場合では少ない計算コストで処理が可能である．一方で，属性数が多くなると抽出されるルールの数が指数関数的に増大する．これに対し，l 値が大きい場合には，非階層クラスタリングを用いた手法を併用することで計算コストを削減可能である．クラスタリングは考慮する属性数によって影響を受けにくいため, この傾向は属性数が大きくなると顕著となる.

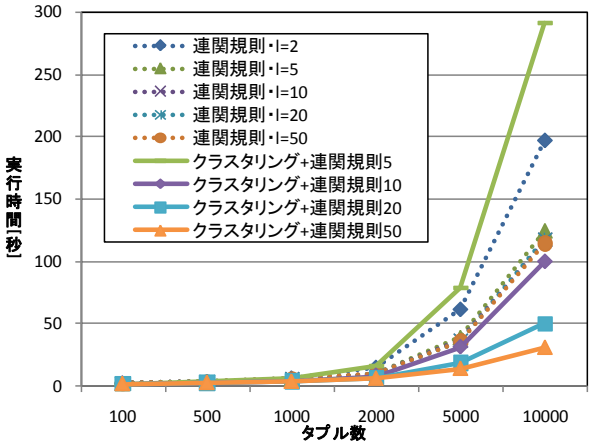

図 3　タプル数の変化による実行時間の傾向

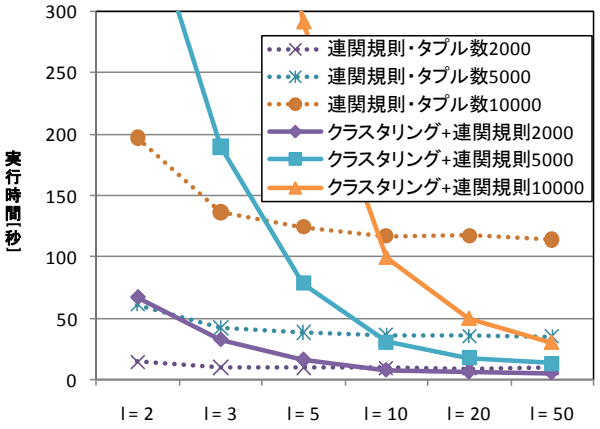

図 4　l 値の変化による実行時間の傾向

7　結論

サービス親和性の高いデータの取得と提供を実現するサービス指向型ルータにおいて，プライバシに配慮した形式でデータを公開可能とする低コストプライバシ保護手法を提案した．本提案を用いることにより，柔軟なプライバシ保護設定の自由度を保ちながら，多量のデータに対してもより少ない実行時間内でプライバシ保護機能を実現可能である.

参考文献

[1] K. Inoue, D. Akashi, H. Nishi. “Semantic Router using Data Stream to Enrich Services”. 3rd International Conference on Internet (CFI), pp, 20-23, Jun, 2008

[2] A. Machanavajjhala, D. kifer, J. Gehrke, M. Venkitasubramaniam. “l-diversity: Privacy beyond k-anonymity”. ACM Transsactions on Knowledge Discovery from Data (TKDD), Vol. 1, No. 1, p. 3, 2007