個人情報開示に対する心理的障壁と漏えい防止技術

近年、個人のプライバシーに関する情報は、公開を拒否する人も多く、個人を特定できないよう、匿名化が必須である。しかし、過度の匿名化によって情報の質が低下し、情報を有効に活用できないという問題も生じている。どうすれば情報を保護しつつ、活用を促すことができるだろうか。

匿名化と統計データ活用のジレンマ

――現在、お２人は新領域融合プロジェクトの１つ、「人間・社会システム」分野において、それぞれどのような研究をされていますか？

越前　私は、これまで大学や研究機関において、閉じて利用することが前提だった個人調査データなどの統計情報を、一定の匿名性を確保することにより、組織の枠を超えて学術的に活用する研究を行っています。

たとえば、ある医療データに、氏名、住所、年齢、病名、投薬に関する情報が記載されていたとします。そのまま公開すれば個人が特定されてしまうので、氏名や住所を削除し、茨城県を関東に、32 歳を 30 ～ 39 歳といった具合に、属性をぼかすことにより、同じ属性をもつ人を複数人存在させ、個人の特定ができないようにするのです。

ところが、このように匿名化を進めれば進めるほど、データの正確さや価値は損なわれていきます。いわば、“データの匿名化の度合い”と“データの学術的有用性”はトレードオフの関係にあるということです。

そうしたなか、2009 年に新統計法が施行され、調査票データの学術利用が可能になりました。それにより、データの匿名化よりも、学術的有用性に重点が置かれるようになった。従来、匿名化されたデータは誰でもアクセスできるという前提で、データの匿名化の度合いを高めることに重点が置かれてきましたが、今後は、匿名化の度合いを弱めて利用しやすくする一方で、利用者によるデータの漏えい防止に重点が置かれることになったというわけです。

小林　確かに、従来、二次分析用に公開される統計データは、使いづらいものが多くありました。例えば、都市と村落の比較分析をしたくても、県レベルの居住地域情報しか付与されていなかったり、個票ではなく集計データしか提供されていなかったり――。そもそも統計法が対象としているデータの中でも、私が扱うような社会調査データというのは、個人を特定できないものがほとんどなので、医療データなどとは異なり、これ以上の個票レベルでの匿名化の必要はないと思います。実際に、統計上、頻度が稀なデータは、トップコーディングといって、欠損値扱いにしていますし、匿名化を強固にする方法論自体は、すでにかなり確立されているのです。

そこで、私のほうは、いかにして人々に個人情報を提供してもらうか、という観点で研究をしています。なかでも現在、学術に限らず期待されているのが、人間行動やコミュニケーションに関するライフログの活用です。しかし、ライフログというのは、それこそプライバシーに関わるデータなので、その提供には抵抗を

越前　功
Isao Echizen
国立情報学研究所
コンテンツ科学研究系准教授

小林哲郎
Tetsuro Kobayashi
国立情報学研究所
情報社会相関研究系准教授

もつ人が多い。そこで、何が心理的障壁になっているのか、いかにして社会にとって有用なデータを提供してもらうことができるのかについて研究を進めています。

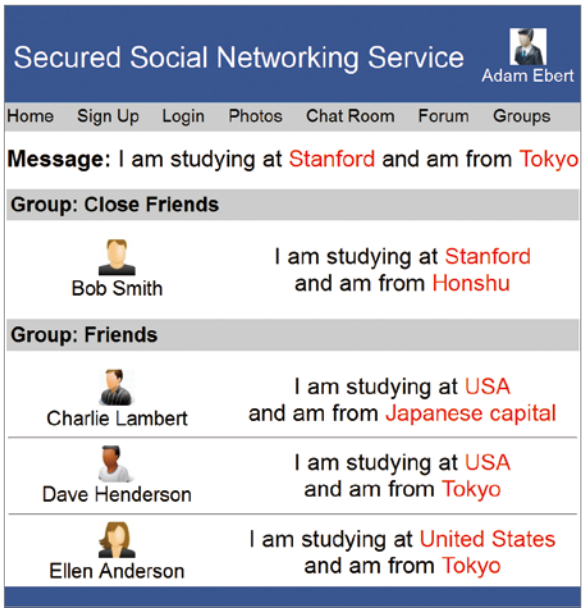
情報漏えい防止の技術と、心理的障壁を押し下げる工夫

――具体的な研究内容をお聞かせください。

越前　匿名化データ向けのフィンガープリント手法といって、データの匿名化プロセスの多義性（同一の匿名性を確保するのに多様な匿名化プロセスが存在すること）を用いて、個々の匿名化プロセスと利用者の識別情報を紐づけすることで、漏えいした匿名化データから、漏えい元を特定できる手法を考案しました（図）。たとえば、生年月日と性別からなるデータであれば、利用者 A には、「1971 年、男性」と書いたものを渡し、利用者 B には「1971 年 8 月 10 日、性別不明」といったデータを渡して、同等の匿名化レベルを確保しながら、利用者ごとに異なる匿名化プロセスで生成したデータセットを用意するのです。その匿名化プロセスと利用者 ID を紐づけすることで、万一漏えいした場合、どの利用者からデータが漏えいしたのか、特定することができるというしくみです。このルールが適用されていることを知ると、利用者は情報の管理に慎重にならざるを得ません。つまり、匿名化プロセスそのものをデータ漏えいの抑止力に使うわけですね。

これを SNS やブログに応用すれば、利用者が属するグループごとに匿名化のレベルを変えながら、利用者ごとに異なる匿名化プロセスで生成したテキストを表示することで、漏えいしたテキストから漏えい元を特定できるようになるのです。

SNSへのフィンガープリント手法の適用
投稿者が投稿したメッセージを、利用者が属するグループごとに異なる匿名化レベルで匿名化する（この例では、投稿者の所属や出身地情報を匿名化している）。グループ内では、同一の匿名化レベルで異なるメッセージを生成する。個々の匿名化プロセスと利用者の識別情報を紐づけしておけば、漏えいしたメッセージから漏えい元が特定できる。



小林　私のほうは、スマートフォンを使用している人を対象に、スマートフォンから得られるライフログ（①位置情報／② Web 閲覧履歴／③通話・SNS・Gmail によるコミュニケーション）の提供を依頼する被験者実験を行いました。その際、3 種類の謝金（① 1000 円／② 5000 円／③ 1 万円）を設け、提供してもらうライフログの種類と金額を組み合わせた条件を設け、各要因が提供率に及ぼす効果について調べました。その結果、やはり謝金が多いほど、情報提供率が上がることがわかった。一方で、いくらお金を積まれても、個人情報は提供したくないという人が、3 割程度いることもわかりました。なかでも、通話記録などコミュニケーションに関わる情報の提供には、内容が記録されないとしても抵抗がある人が多いようです。逆に、GPS データは比較的ハードルが低いようです。位置情報の提供にはリスクを伴うはずですが、そのことはあまり認識されていないのかもしれません。翻って、GPS データのように、比較的ハードルの低い情報から徐々に間口を広げ、心理的障壁を取り払い、提供してもらう情報の種類を増やしていくこともできるのではないかと考えています。

個人情報の提供には高い障壁があり、またその漏えいはあってはなりませんが、一方で、皆が提供すれば、それだけ全体としての情報の質が向上し、「公共財」として利用価値が高まっていくと考えられます。従って、情報提供を促進するインセンティブの設計については、今後いっそうの研究が必要だと思います。

越前　同感です。データの匿名性と有用性という、相反する要素を扱うのは難しいことですが、それをうまくバランスさせる技術を開発することで、情報の有効活用に貢献できるよう、今後も研究に取り組んでいきたいと思います。

（取材・構成　田井中麻都佳）