3C2-OS-13b-6

プライバシーを含む Linked Data の秘匿分析技術

An approach to privacy-preserving processing of anonymized Linked Data

津田 宏
Hiroshi Tsuda

伊藤 孝一
Koichi Ito

株式会社富士通研究所 ソフトウェアシステム研究所
Software Systems Laboratory, Fujitsu Laboratories Ltd.

This paper describes an information-centric approach to privacy-preserving processing of Linked Data. We applied our information masking technology, which enables secure inter-cloud data integration, to RDF so as to realize access control in RDF mining. This paper also presents a prototype of KnowWho search from anonymized RDF.

1. はじめに

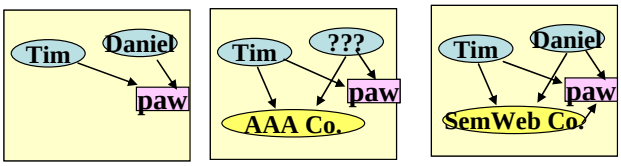
Linked Data を構成する RDF や OWL は、LOD(Linked Open Data)としてだけでなく、社内の情報統合・活用の基盤としても活用可能である。ただし、LOD と異なり、社内では権限に応じた情報のアクセス制御が必要となる。そこで問題となるのは、RDF 自体にはアクセス制御の仕組みが内在されていない点である。

一方、クラウドのように第三者に処理を委託する形態の発展に伴い、機密情報や個人情報をクラウドに預けて処理するための、秘匿化(匿名化)技術の研究が進められている。データに含まれる個人情報をあいまい化したり、暗号化したまま一部の処理が可能な手法で安全化し、クラウドでは秘匿性を保持したまま処理を行うことができる。本論文では、我々が開発したクラウド向けの情報ゲートウェイ技術を Linked Data に応用した試みについて述べる。

2. Linked Data の社内活用とその課題

我々が以前開発したビジネス情報ナビゲータは、社内の報告書やグループウェアなど異種の情報源から、情報抽出技術により RDF を自動生成し、マイニング・見える化を行う枠組みである[津田 2008]。人やミーティングなどの共通オントロジー(OKAR)により RDF 化することで異種情報も意味的に統合し、人脈関係などをマイニングすることが可能である。特に、人とスキル、組織にまたがる複雑な関係もグラフ形式で見える化することで、利用者も把握しやすくなる。本技術の一部は銀行業務でも使われている。

ただし、社内データの活用においては、特定の役職以上の人だけ特定の情報にアクセス可能など、トリプルをベースとしたアクセス制御が必要となる。例えば、下図で Tim と Daniel が paw プロジェクトと SemWeb Co.に関係している、というような情報も、利用者権限に応じて、ノードを消去したり、企業名や人名を匿名にするなどが必要となる。

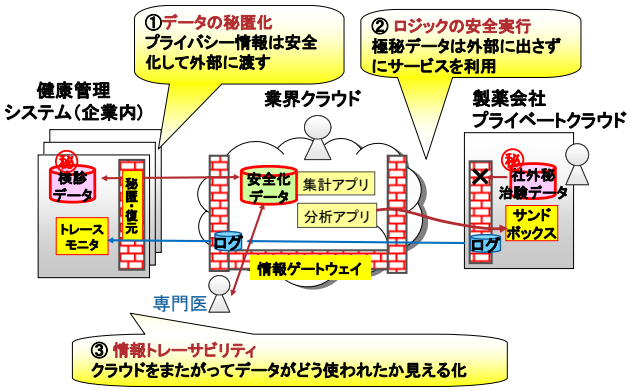


連絡先:津田宏, 富士通研究所 ソフトウェアシステム研究所, htsuda@jp.fujitsu.com

RDF や SPARQL にはこうしたアクセス制御は内在されていないため、例えば RAP[Reddivari 2005] のように、RDF ストレージに外付けでアクセス制御機能を付加する試みがある。ただし、RDF ストレージには実データが入っているため、第三者に委託する場合には新たな問題が生じる。例えば、社内であっても、顧客情報を他の部署に提供できない場合があるとか、クラウドにこうした個人情報を提供するのに不安などである。

3. 情報ゲートウェイ

我々は、複数のクラウドを融合して安全に情報を連携し、活用するための技術として、情報ゲートウェイを開発している[伊藤 2011]。これは、下図にあるように、情報のフィルタリング、トレーサビリティ、サンドボックス、秘匿・復元機能を持ったゲートウェイ(Proxy)であり、オンプレミスやクラウドに配置することで、クラウド間で機密情報や個人情報をフィルタリング・秘匿化してやりとりすることが可能になる。



3.1 秘匿化

利用者がクラウドにデータを預ける不安は根強く、その解決として各種のプライバシ保護マイニング(PPDM)技術が開発されている。我々は秘匿化として、以下のような技術を開発している。いずれも、クラウドに送信する手元(例えば企業内やクライアント)で秘匿し、クラウドでは秘匿したままの処理を行い、結果を手元で復元、ということで安全性を確保している。

- 情報を落とす: 墨塗り(マスキング)、機密部分削除など
- 情報を丸める: 乱数付与(摂動)、ブラーリング(平均値に置き換え)、集合匿名化(k-匿名)など
- 情報を置き換える: ハッシュ、トークン化、暗号化など

- 秘匿処理: 準同型暗号、秘匿集計など秘匿したまま特定の処理が可能

3.2 秘匿集計

ここでは秘匿集計技術[伊藤 2011]について概説する。例えば地域別の月別感染者数や、支店別の月別売上などの表形式の機密情報を対象としている。各企業としては特定の地域に問題が多いという情報を他社に知らせたくない一方、クラウドではこうした情報を各社から集計することで、二次情報の利活用が可能となる。我々の秘匿集計技術は、この解決手法の一つで、以下の2つの技術からなる。

1) 住所情報のユーザ権限に応じた暗号化

「神奈川県/川崎市/中原区」のような階層を持ったデータに対して、ユーザの権限に応じて復号できるレベルを制御できる暗号化である。s1/s2/s3 というデータに対して、R1(s1+s2+s3 のビット長),R2(s2+s3 のビット長),R3(s3 のビット長)という乱数を生成し、R1⊕ R2⊕ R3(⊕ は XOR 演算)にて暗号化する(実装にあたっては、AES カウンターモード等で可能である)。最高レベルの人には R1,R2,R3 を、中レベルの人には R1,R2 を、低レベルの人には R1 を鍵として配ることで、低レベルの人は「神奈川県」までしか復号することができないなどの制御を行うことが可能になる。

2) 表データのスクランブル集計

クラウドで鍵を持たずに、複数の表の集計を秘匿したまま行い、かつ縦横の正しい合計値を得ることができる。これには、縦横の和がゼロになる行列 RM を、RM=RM1+RM2+…+RMn と分割して、n 人のデータ提供者に鍵として提供する。各データ提供者は、元データ表に RMi を加算したものをクラウドに預けることで、クラウドでの集計結果において、縦横の和は元の合計値と等しくなる。

4. 秘匿 KnowWho 検索への適用

4.1 Linked Data の秘匿化(仮名化)

Linked Data(RDF)の秘匿化として、ここでは Subject, Object のみを対象と考える。秘匿化の対象としては、URI およびリテラルになる。本論文で想定する応用については、秘匿化された状態で同一性が判定できないといけないため、決定暗号またはハッシュ、ID 置換などの技術が必要である。

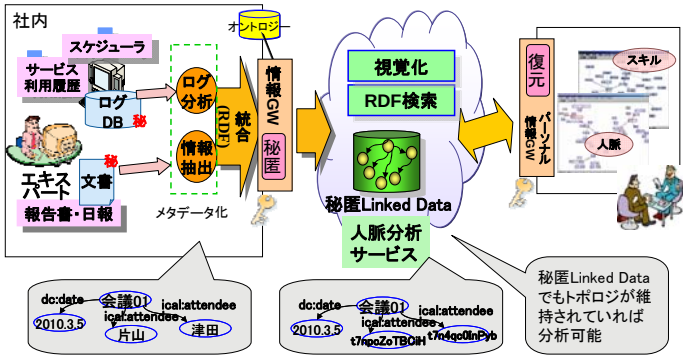
URI は階層的な構造を持っているため、前述の住所情報の秘匿と同様な手法が適用可能である。例えば、

http://www.fujitsu.com/people/hiroshi_tsuda

という URI に対して、利用者のレベルに応じて、人を特定したり、または所属までしか特定できないなどの制御が可能になる。

リテラルについては、通常の暗号化、または前述の表形式の数値のようなプライバシ保護技術が利用可能である。ただし、文字列の部分一致などはできなくなる。

一般に匿名化を行う場合、どの部分が個人・機密情報かはシステムやデータ毎に異なり、実際には人手で一つ一つ指定するなど工数がかかることが多い。Linked Data の場合、オントロジーを活用することで自動的に匿名化も可能と考えられる。例えば dc:creator やその下位プロパティの値は人名なので匿名化するなどである。

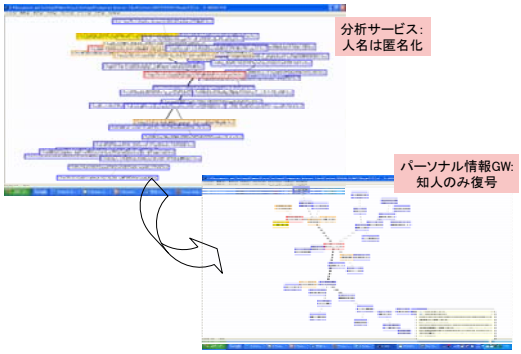


4.2 適用例

情報ゲートウェイによる Linked Data の匿名処理システムの概要を左下図に示す。人脈分析サービスには秘匿化された Linked Data を与え、その状態でマイニングを行う。

マイニングされた結果は、情報ゲートウェイを通してアクセスすることが可能であり、ユーザはインタラクティブに分析を行うことができる。

下図は試作したクライアント版情報ゲートウェイによる実行例である。サービス側では人名(ノード)は秘匿化されているが、ユーザ側では知人は復号されている。同じ分析結果から、ユーザに応じて見える情報をリアルタイムに制御する等が可能となる。



5. おわりに

本稿では、社内データの統合活用に Linked Data(RDF)を利用する際に必要となるアクセス制御を、暗号技術を使った秘匿(仮名)Linked Data により実現した試みについて述べた。また、個人情報の検索・マイニングのプロトタイプによる適用例も示した。従来の外付けでアクセス制御をする技術と異なり、データのみでアクセス制御を行うため、分析処理には手を入れる必要がなく、クラウドのような環境でも適用できる手法である。

本稿で述べた匿名化手法は、仮名 RDF というべきで、本人の非到達性を満たしているが、KnowWho という応用上リンク不能性は満たしていない[折田 2009]。応用によってはさらなる匿名化技術が必要である。また、実際には社内データと LOD のような外部データとの組み合わせによる分析も必要となる。そのためには、例えば外部への検索クエリから情報が漏れないようなプライバシ保護の仕組みも必要となる。

今後、EU データ保護指令の強化など、プライベートデータを確実に保護しつつ利用することが求められており、Linked Data のような意味的なデータ記述の枠組みと、プライバシ保護技術との連携は今後ますます重要になると考えている。

参考文献

[Reddivari 2005] Pavan Reddivari etc. , "Policy based Access Control for a RDF Store", Proceedings of the Policy Management for the Web Workshop, WWW Conf. 2005.

[津田 2008] 津田, ビジネスに生かすメタデータの統合、「見える化」技術, INTAP セマンティック Web コンファレンス 2008.

[伊藤 2011] 伊藤,片山,牛田,高,小櫻,津田, クラウドにおけるデータ秘匿化および追跡技術, 電子情報通信学会IN研究会, 2011.1

[折田 2009] Web 上の人物および行為の信頼性評価, 人工知能学会誌, Vol.24, No.4, pp.527-534, 2009.