개인정보보호를 위한 RDF의 확장

김윤삼, 조은선
충남대학교 컴퓨터공학과
e-mail:{bijak,escho}@cnu.ac.kr

RDF Extension for Privacy Protection

Yun-Sam Kim, Eun-Sun Cho
Dept of Computer Engineering, Chung-Nam University

요 약

유비쿼터스 시스템이 발전함으로 많은 사람들이 시스템을 사용함으로 인하여 정보보호에 대한 중요성이 증대되고 있다. 본 논문에서는 유비쿼터스 시스템에서 데이터를 표현하기 위하여 널리 사용되는 RDF/S가 정보보호에 대하여 취약함에 주목하고 RDF/S에 대한 정보보호를 위한 RDF/S 스키마와 추론에 필요한 규칙을 소개한다.

1. 서론\

최근 유비쿼터스 시스템이 발전함에 따라 시스템에 저장되는 개인정보의 종류와 양이 급격하게 늘고 있다. 그러나 이러한 환경에 대하여 외부자의 해킹이나 내부자의 자의적인 개인정보 접근에 의한 개인정보가 유출되는 상황이 발생함으로 인하여 개인정보의 중대성에 대한 인식이 증대되고 있는 상황이다.

본 논문에서는 유비쿼터스 환경에서 데이터를 표현하기 위하여 사용되는 RDF(Resource Description Framework) 질의에 대한 개인정보보호를 구현하기 위한 RDF 스키마와 규칙을 제안한다. 2장에서는 현재 RDF에서의 개인정보의 기술과 개인정보보호를 위해 정의한 RDF 스키마와 추가적인 규칙, 그리고 이를 이용한 추론에 대하여 설명하며, 3장에서는 확장된 RDF 데이터에 대한 질의에 대하여 정보보호기법을 어떻게 적용하게 되는지에 대하여 설명한다. 4장에서는 관련 연구로써 데이터 테이블과 데이터베이스에서 연구되어진 개인정보보호기법에 대하여 설명한다.

2. 개인정보보호를 위한 RDF의 확장

2.1 RDF을 이용한 개인정보의 표현

기존의 데이터베이스가 고정된 형태의 데이터만 표현 가능했던 것에 비하여 RDF는 XML을 이용하여 복잡한 형태의 데이터를 표현하는 것이 가능하다[1]. 그러나 이러한 RDF는 그림 1과 같이 개인정보를 표현할 경우 질의에 대하여 추상화된 결과를 보여주지 못한다. 따라서 사용자의 질의에 대하여 해당하는 결과의 전부를 보여주거나 전부를 보여주지 않는 All or Nothing의 결과만을 보여주게 된다. 즉, 그림 1에서 'YunSam_Kim'의 위치에 대한 질의를 할 경우 'Room Number 336'을 질의 결과로 보여주거나 질의를 거절하는 둘 중 하나를 택하여야 한다. 'Room Number 336' 대신 'Third Floor'와 같은 추상화된 결과를 보여줄 수 없다.

```
<rdfs:Description rdf:http://plas.cnu.ac.kr/rdf:rdfsData#YunSam_Kim>
        <ex:location>Room_Number_336</ex:location>
</rdfs:Description>
<rdfs:Description rdf:http://plas.cnu.ac.kr/rdf:rdfsData#JoonYoung_Paik>
        <ex:location>Room_Number_101</ex:location>
</rdfs:Description>
```

(그림 1) 개인정보보호가 않되는 RDF 데이터

따라서 우리는 필요에 따라 쿼리의 결과를 좀 더 추상화하여 보여주기 위한 기법을 제안하며, 이를 위한 확장된 RDF 스키마와 추론 규칙에 대하여 기술한다.

2.2 확장된 RDF 스키마

RDF가 개인정보를 보호하기 위하여서는 기존에 제공되는 RDF만으로는 불충분하다. 따라서 개인정보보호를 위한 추가적인 RDF 스키마를 정의해야 한다. 이를 위한 확장된 RDF 스키마는 그림 3과 같다.

Anonymizer는 k-Anonymity[2]의 Domain Generalization Hierarchy(DGH)와 같은 추상화 정책에 따라서 나타날 수 있는 개인정보의 구조를 나타낸다. 각각의 Anonymizer는 추상화 가능한 정보(Generalizable Information)들과 추상화 불가능한 정보(Ungeneralizable Information)들을 가지고 있으며, Anonymizer는 subAnonymizerOf의 Relation을 이용하여 각 Anonymizer 간의 계층관계를 표현낸다. 이렇게 subAnonymizerOf로 정의된 Anonymizer들은 추론에 의하여 최종적으로 트리 형태의 구조로 나타내어지며, 상위 계층의 Anonymizer는 하위 계층의 Anonymizer의 추상화 가능한 정보보다 좀 더 추상화된 정보들을 가지게 된다.

추상화 가능한 정보는 보통 사용자의 질의에 대한 결과

```
<rdfs:Class rdf:ID="Anonymizer">
  <rdfs:label xml:lang="en">Anonymizer</rdfs:label>
  <rdfs:comment>This Presents the set of Anonymizer</rdfs:comment>
</rdfs:Class>

<rdfs:Resource rdf:ID="GeneralizableInformation">
  <rdfs:label xml:lang="en">Quasi Identifier</rdfs:label>
</rdfs:Resource>

<rdfs:Resource rdf:ID="UngeneralizableInformation">
  <rdfs:label xml:lang="en">Secure Information</rdfs:label>
</rdfs:Resource>

<rdf:Property>
  <rdfs:label>subAnonymizerOf</rdfs:label>
  <rdfs:comment>The subject is a subclass of a Anonymizer</rdfs:comment>
  <rdfs:range rdf:resource="#Anonymizer"/>
  <rdfs:domain rdf:resource="#Anonymizer"/>
</rdf:Property>

<rdf:Property>
  <rdfs:label>GeneralizableInformationOf</rdfs:label>
  <rdfs:comment>The subject is a subset of a Anonymizer</rdfs:comment>
  <rdfs:range rdf:resource="#GeneralizableInformation"/>
  <rdfs:domain rdf:resource="#Anonymizer"/>
</rdf:Property>

<rdf:Property>
  <rdfs:label>UngeneralizableInformationOf</rdfs:label>
  <rdfs:comment>The subject is a subset of a Anonymizer</rdfs:comment>
  <rdfs:range rdf:resource="#UngeneralizableInformation"/>
  <rdfs:domain rdfs:resource="#Anonymizer"/>
</rdf:Property>
```

(그림 2) 확장된 RDF 스키마

이며, 질의에 대하여 정확한 결과를 보여주지 않기를 원하는 정보를 나타낸다. 즉, 'YunSam_Kim이 어디있는가?'라는 질의에 대하여 정보의 소유자는 그 결과가 'Room_Number_336'이라는 것에 대하여 이를 그대로 보여주기를 원치 않은 경우가 생길 수 있다. 이러한 경우 필요에 따라서 그 결과를 'Third_Floor' 또는 'Engineering_Building'과 같이 좀 더 추상화된 결과를 질의에 대한 응답을 주게 된다. 이러한 추상화된 결과는 계층구조를 가지게 되며, Anonymizer의 관계에 따라서 추상화 가능한 정보는 좀 더 추상화된 값을 가지게 된다. 즉, 트리 구조에서 상위의 Anonymizer는 하위의 Anonymizer에 비하여 더욱 추상화된 추 상화 가능한 결과를 갖게된다. 이러한 추상화 가능한 정보는 GeneralizableInformatonOf의 관계를 이용하여 Anonymizer에 연결된다.

추상화 불가능한 정보는 추상화를 할 경우 그 의미를 잃어버리거나 사용자의 질의에 사용되는 등 온전한 데이터의 형태를 유지해야 할 필요가 있는 정보를 의미한다. 즉, 그림 1에서의 'YunSam_Kim'과 같은 정보가 추상화 불가능한 정보에 속한다. 이 추상화 불가능한 정보는 추론에 의하여 상위 Anonymizer의 추상화 불가능한 정보에도 속하게 된다. 이를 통하여 'Kim이 어디있는가?'라는 질의에 대하여 'Room_Number_336', 'Engineering_Building'과 같은 결과를 보여줄 수 있다. 이러한 추상화 불가능한 정보는 UngeneralizableInformationOf의 관계 이용하여 Anonymizer와 연결된다.

2.2 RDF 추론

확장된 RDF 스키마 그대로는 정보보호를 만족할 수 없다. 하위 Anonymizer에 속해 있는 추상화 불가능한 정보는 상위 Anonymizer에도 존재하여야 사용자의 질의에 대한 응답을 추상화 시킬 수 있기 때문이다. 이를 위하여

<표 1> 추가된 RDF 규칙

규칙	규칙의정의
rule1	(?A subAnonymizerOf ?B) (?B subAnonymizerOf ?C) -> (?A subAnonymizerOf ?C)
rule2	(?A UngeneralizableInformationOf ?B) (?B subAnonymizerOf ?C) -> (?A UngeneralizableInformationOf ?C)

확장된 RDF 스키마는 RDF/S에서 제공하는 추론 규칙 이외에 추가적으로 정보보호를 만족시키기 위한 추론 규칙을 가져야 하며. 그 규칙은 표 1과 같다.

규칙 1은 확장된 스키마 subAnonymizerOf의 이행성을 정의한다. 각각의 Anonymizer들은 프로퍼티 subAnonymizerOf를 통해 트리 형태의 구조를 띄게 된다.

그러나 RDF 스키마는 하나의 Anonymizer에 대하여 상위 Anonymizer에 대한 관계만을 정의한다. 따라서 추론 규칙을 추가하지 않을 경우 Anonymizer의 조상(Ancestor)와의 관계를 정의 할 수 없다. 이러한 모든 Anonymizer들의 관계를 정의하기 위하여 규칙 1을 사용한다.

규칙 2는 Ungeneralizable Information의 확산을 정의한다. 2.2에서 설명하였듯이 추상화 가능한 정보는 상위 Anonymizer에 대하여 정보가 사용자에 의하여 명시적으로 추상화되어 나타나짐에 비하여 추상화 불가능한 정보는 하위의 정보가 상위의 정보에 대하여 그대로 나타나야 한다. 그러나 관리자 또는 시스템이 나타내는 추상화 불가능한 정보는 최하위의 Anonymizer에만 연결이 된다. 따라서 모든 추상화 불가능한 정보는 상위 Anonymizer로의 확산을 하여야 하며, 이는 추론 규칙 2를 통하여 이루어진다..

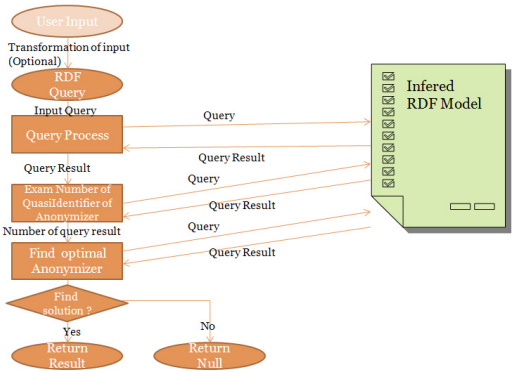
이렇게 확장된 RDF 스키마와 추가된 RDF 규칙을 이용하여 RDF로 표현된 데이터는 Sesame[2], Jena[3]과 같은 추론엔진을 이용하여 개인정보보호를 만족할 수 있도록 추론된다. 그림 2는 추론이 이루어진 후의 RDF 데이터의 일부분을 나타낸 것이다. 그림 3의 'JoonYoung_Paik'은 기존의 RDF 데이터에서는 '_10'의 Anonymizer에 연결된 추상화 불가능한 정보이며, '_9'와 '_1'은 추론에 의하여 연결된 것이다.

```
<rdf:Description rdf:about="http://plas.cnu.ac.kr/rdf/rdfsData#JoonYoung_Paik">
    <ex:UngeneralizableInformationOf rdf:resource="http://plas.cnu.ac.kr/rdf/rdfsData#_9"/>
    <ex:UngeneralizableInformationOf rdf:resource="http://plas.cnu.ac.kr/rdf/rdfsData#_1"/>
    <ex:UngeneralizableInformationOf rdf:resource="http://plas.cnu.ac.kr/rdf/rdfsData#_10"/>
    <rdf:type rdf:resource="http://plas.cnu.ac.kr/rdf/rdfsSchema.rdf#GeneralizableInformation"/>
    <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
</rdf:Description>
```

(그림 3) 확장된 RDF 데이터

3. 개인정보보호를 위한 사용자 질의에 대한 처리

사용자가 개인정보보호를 만족하는 RDF 데이터에 대하여 질의를 할 경우 조건을 만족하는 결과는 하나 이상이 될 수 있다. 즉, 'Kim이 어디있는가?'라는 질의에 대하



(그림 4) 사용자 질의의 처리 순서

여 '336호', '3층', '공과대학'이라는 세 가지의 결과가 나오게 되며, 시스템은 질의에 대하여 문제가 없을 경우 세 가지 결과 중 하나를 사용자에게 알려줘야 한다. 이러한 질의에 대하여 시스템은 미리 정해놓은 값에 따라 각 Anonymizer의 추상화 불가능한 정보의 수를 조사하여 정해놓은 값보다 큰 Anonymizer 중 트리 구조에서 사장 하위의 Anonymizer의 추상화 가능한 정보를 결과 값으로 전달하게 된다. 그 과정은 그림 3과 같다.

먼저 사용자가 시스템에 맞게 작성한 질의는 표 2의 첫 번째 질의와 같이 RDF에 맞는 질의로 변환된다. 이렇게 변환된 질의를 이용하여 시스템은 추론된 RDF 데이터를 이용하여 추론을 한다. 그림 2의 경우에는 그 결과로 'Room_Number_336','Third_Floor','Engineering_Building' 등을 추상화 가능한 정보로 갖는 Anonymizer들이 결과로 나오게 된다. 이렇게 나온 결과에 대하여 표 2의 두 번째 질의의 형식으로 각각의 추상화 불가능한 정보의 개수를 파악하게 된다. 이에 대하여 미리 정의된 값 k보다 크거나 같은 수 중 가장 작은 값을 갖는 Anonymizer를 찾는다. k가 3이며, 'Room Number 336'의 Anonymizer가 가지고 있는 추상화 불가능한 정보의 값은 2, 'Third Floor'에 대한 값은 3, 'Engineering Building'에 대한 값이 4이라 할 경우 'Room Number 336'의 2는 3보다 작으므로 질의에 대한 결과가 되지 못한다. 3보다 큰 경우로는 3과 4가 있으며 이 중 3이 작으므로 원하는 결과를 가지고 있는 것은 3층을 추상화 가능한 정보의 값으로 갖는

<표 2> 사용자 질의에 의하여 생성되는 시스템 질의

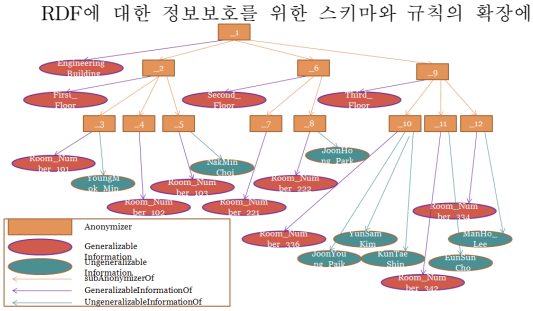
1	SELECT ?z WHERE {QueryData> <GeneralizableInformationOf> ?z. }
2	SELECT ?x WHERE {?x <UngeneralizableInformationOf> <1의 결과값>.}
3	SELECT ?x WHERE {<첫번째 Anonymizer> ?x <두번째 Anonymizer>.}
4	SELECT ?x WHERE {?x <GeneralizableInformationOf> <Result Anonymizer>.}

Anonymizer가 된다. 이렇게 최적의 Anonymizer를 찾은 경우 표 2의 네 번째 질의를 이용하여 사용자가 원하는 결과 'Third_Floor'를 돌려주게 된다.

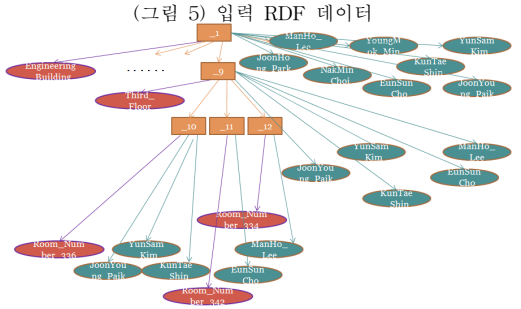
만약 'Engineering Building'의 Anonymizer가 가지고 있는 추상화 불가능한 정보의 개수 또한 4인 경우에는 표 2의 세 번째 질의를 이용하여 각 Anonymizer의 관계를 조사하여 좀 더 낮은 레벨의 Anonymizer인 'Third Floor'의 Anonymizer를 최적의 결과로 찾게 된다. 이 때 상하 관계를 파악하기 위하여 두 Anonymizer 사이에 프로퍼티가 존재하는지 확인하여 존재하면 첫 번째 Anonymizer는 두 번째 Anonymizer에 비하여 하위 Anonymizer라고 판단한다. 만약 k보다 큰 Anonymizer가 없을 경우 시스템은 사용자의 질의에 대하여 거절하게 된다. 이를 통해서 시스템은 질의에 대한 신뢰성을 높일 수 있다.

4. 실험

RDF에 대한 정보보호를 위한 스키마와 규칙의 확장에



(그림 5) 입력 RDF 데이터



(그림 6) 추론된 RDF 데이터의 일부

대한 실험은 Jena와 Java를 이용하여 진행되었으며, 이를 위한 RDF 데이터의 트리구조는 그림 4와 같다. 이를 Jena 추론 엔진으로 추론을 진행하면 RDF 데이터가 확장되며, 그림 5는 그 중 일부분을 나타낸 것이다.

실험에 의하여 사용된 질의는 그림 6과 같으며, k 값을 2에서 9까지 증가시켜 실험하였으며, 그 결과는 표 3과 같다.

```
SELECT ?location WHERE
{<http://plas.cnu.ac.kr/rdf/rdfsData#YunSam_Kim>
 <http://plas.cnu.ac.kr/rdf/rdfsSchema.rdf#UngeneralizableInformationOf>
 ?location.}
```

(그림 7) 실험에 사용된 질의

<표 3> 실험 결과

k	result	number of query
2	Room_Number_336	5
3	Room_Number_336	5
4	Third_Floor	5
5	Third_Floor	5
6	Engineering_Building	5
7	Engineering_Building	5
8	Engineering_Building	5
9	null	5

4. 관련 연구

4.1 k-Anonymity

P. Samarati에 의하여 소개된 k-Anonymity는 데이터 테이블 또는 데이터의 추가 및 삭제가 발생하지 않는 고정적인 데이터에 대한 개인정보보호 기법이다. 추상화시켜야 하는 정보(Quasi Identifier)를 Generalization과 Suppression을 이용하여 같은 값을 k개 이상 만들어 개인정보(private data)를 알 수 없도록 하는 기법이다. 이러한 k-Anonymity 기법의 보완으로 l-Diversity[5], t-Closeness[6]와 같은 기법이 소개되었다. 그러나 이러한 기법들은 자료의 추가와 삭제를 고려하지 않았으며, 질의 또한 고려하지 않아 이를 복잡한 형태의 자료 표현에 그대로 적용할 수는 없다.

4.2 Database에서의 k-Anonymity

데이터베이스에서의 개인정보보호는 데이터의 삽입과 삭제가 발생할 수 있으며, 질의에 의하여 일부분의 데이터만을 추출할 수 있다는 특징에 의하여 고정적인 데이터의 k-Anonymity등을 적용하지 않는다. 대신 발생할 수 있는 모든 질의에 대하여 데이터베이스가 k-Anonymity를 만족할 수 있는지에 대하여 검사한다[7]

이러한 기법은 Generalization을 사용하지 않기 때문에 데이터베이스에 함수 의존성(Functional Dependency)가 있는 경우 모든 질의에 대하여 k-Anonymity를 만족할 수 없는 문제가 생긴다.

5. 결론

개인정보보호는 현재 부각되고 있는 중요한 이슈중 하나이다. 본 논문으로 RDF에 이러한 개인정보보호를 적용하기 위한 RDF 스키마와 규칙을 정의하였다. 그러나 이러한 기법은 질의의 횟수가 늘어나는 문제점을 가지고 있어 시스템의 성능에 좋지 않은 영향을 미칠 수 있다. 따라서 추후 RDF 데이터에 대한 최적화 및 정보의 저장을 이용한 질의의 횟수를 줄일 계획이다.

참고문헌

[1] W3C, "Resource Description Framework", http://www.w3.org/RDF/

[2] P. Samarati and L. Sweeney "Generalizing data to provide anonymity when disclosing information", Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, 1998

[3] ADUNA Open Source project, http://www.openrdf.org/

[4] Jena - A Semantic Web Framework for Java, http://jena.sourceforge.net

[5] Ashwin Machanavajjhala, Johannes Gehrke, and Daniel Kifer "l-Diversity: Privacy Beyond k-Anonymity', 22nd IEEE International Conference on Data Engineering, 2006.

[6] Nighui Li, Tiancheng Li, and Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", 23th IEEE International Conference on Data Engineering, 2007.

[7] Chao Yao, X. Sean Wang, and Sushil Jajodia "Checking for k-Anonymity Violations by Views", Proceedings of the 31st international conference on Very large databases, 2005.