Journal of The Korea Society of Computer and Information
Vol. 22 No. 7, pp. 55-62, July 2017
www.ksci.re.kr
https://doi.org/10.9708/jksci.2017.22.07.055

A Study on Performing Join Queries over K-anonymous Tables

Dae-Ho Kim*, Jong Wook Kim**

Abstract

Recently, there has been an increasing need for the sharing of microdata containing information regarding an individual entity. As microdata usually contains sensitive information on an individual, releasing it directly for public use may violate existing privacy requirements. Thus, to avoid the privacy problems that occur through the release of microdata for public use, extensive studies have been conducted in the area of privacy-preserving data publishing (PPDP). The k-anonymity algorithm, which is the most popular method, guarantees that, for each record, there are at least k-1 other records included in the released data that have the same values for a set of quasi-identifier attributes. Given an original table, the corresponding k-anonymous table is obtained by generalizing each record in the table into an indistinguishable group, called the equivalent class, by replacing the specific values of the quasi-identifier attributes with more general values. However, query processing over the anonymized data is a very challenging task, due to generalized attribute values. In particular, the problem becomes more challenging with an equi-join query (which is the most common type of query in data analysis tasks) over k-anonymous tables, since with the generalized attribute values, it is hard to determine whether two records can be joinable. Thus, to address this challenge, in this paper, we develop a novel scheme that is able to effectively perform an equi-join between k-anonymous tables. The experiment results show that, through the proposed method, significant gains in accuracy over using a naive scheme can be achieved.

▸Keyword: Privacy-preserving data publishing, K-anonymity, Generalization, Join query

I. Introduction

IoT 시대의 도래 등으로 인해, 현재 많은 산업분야에서 빅데이터가 폭발적으로 생성되고 있다. 최근 수년간 빅데이터에 대한 분석이 큰 가치를 창출할 것으로 주목 받아왔으며, 현재 관련 기술 또한 활발히 개발되고 있다. 빅데이터 분석에 대한 관심이 높아지고 있는 만큼 빅데이터를 공유하고 활용하는 것에 대한 관심 또한 높아지고 있다. 이러한 흐름에 발 맞춰 정부는 교통, 인구 등 다양한 공공 빅데이터를 개방하였고 사용자들은 이 빅데이터를 활용하여 새로운 가치를 창출해오고 있다.

하지만 동전의 양면처럼 빅데이터 공유와 활용은 다른 방향으로 악용이 될 수 있다. 동영상 스트리밍 회사 넷플릭스의 경우 영화 추천 알고리즘의 정확성을 높이기 위해 개최한 경연대회에서 50만명 이용자의 영화 평가 데이터 1억 건을 비식별화한 후 공개하였다. 그러나 텍사스 대학 연구팀은 온라인 영화 전문 사이트에 공개된 영화 평가 데이터를 이용하여 비식별화된 넷플릭스 데이터로부터 개인의 민감한 정보를 재식별화 할 수 있었다 [1]. 이처럼 빅데이터의 공유와 활용이 증가함에 따라 개인정보 유출 위험성도 증가하므로, 데이터 속에 존재하는 개인들의 민감한 정보 유출을 방지하기 위한 노력이 요구된다.

최근 들어 개인 정보가 포함된 데이터를 배포하기 위해 개인의 프라이버시를 보호하면서 데이터를 최소한으로 변형하는 프

• First Author: Dae-Ho Kim, Corresponding Author: Jong Wook Kim
*Dae-Ho Kim (rlaeogh222@gmail.com), Dept. of Computer Science, Sangmyung University
**Jong Wook Kim (jkim@smu.ac.kr), Dept. of Computer Science, Sangmyung University
• Received: 2017. 04. 28, Revised: 2017. 05. 18, Accepted: 2017. 06. 18.
• This research was supported by a 2016 Research Grant from Sangmyung University.

라이버시 보호 데이터 배포 (Privacy-Preserving Data Publishing, PPDP)가 활발히 연구되어 왔다 [2,3,4]. 프라이버시 보호 데이터 배포는 프라이버시 모델에 따라 원본 데이터를 변형하는 방식이 다르며, 현재까지 다양한 프라이버시 모델이 연구되어 왔다 (예, k-익명성[5,6], l-다양성[7], t-근접성[8]). 가장 대표적인 k-익명성은 적어도 k개의 데이터가 준식별자 속성에 대해 같은 값을 가지도록 데이터를 변형하는 방법이다 [5,6]. 또한 l-다양성은 k-익명성을 보장함과 동시에 데이터 집합에서 l개의 민감한 정보를 분포시켜 데이터를 비식별화 하는 방법이다 [7]. 이처럼 현재 다양한 비식별화 기술을 적용함으로써 빅데이터의 공유와 활용에 있어 개인 정보 유출을 방지하고 있다.

실제 응용프로그램 환경에서 데이터베이스는 여러 개의 테이블들과 테이블들 사이의 관계들로 구성되어 있다. 이러한 경우, 데이터 사용자들은 데이터 분석 목적에 따라 테이블 간의 조인 연산을 요구하는 다양한 형태의 조인 질의를 사용하게 된다. 그러나 프라이버시 보호 데이터 배포 기법에 의해 익명화된 테이블의 경우, 원본 데이터의 변형으로 인하여 기존 조인 기법을 통하여 정확한 조인 질의 결과를 얻을 수 없다는 문제점이 있다. 그러므로 본 논문에서는 익명화된 데이블 간의 조인 연산을 수행하기 위한 방법을 연구한다. 또한, 프라이버시 보호 데이터 배포에서 데이터를 변형하는 방식이 조인 질의 수행 결과에 어떠한 영향을 미치는지 비교 분석 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서 배경 지식과 본 논문에서 다루는 문제를 정의하고, 3장에서 익명화된 테이블 간의 조인 질의를 수행하기 위한 방법을 설명한다. 4장에서 제안 방법의 성능 평가를 수행한 후, 5장에서 결론을 맺는다.

II. Background and Problem Definition

1. K-anonymity and Generalization

하나의 데이터 속에는 수많은 정보들이 담겨 있다. 이 데이터 속에서 특정한 개인을 식별할 수 있는 속성을 '식별자(identifier)' 라고 한다. 식별자의 예로는 주민등록번호, 휴대폰 번호 등이 있으며, 식별자는 개인 정보 유출을 직접적으로 유발한다. 그래서 데이터 공유에는 식별자들을 삭제하고 공유하는 것이 원칙이다. 하지만 데이터 내에서 식별자를 제거해도, 넷플릭스의 경우처럼 다른 데이터를 결합하게 되면 개인을 식별할 가능성이 있는 속성이 있다. 이렇게 단일 데이터만으로는 개인을 식별할 수 없지만 다른 데이터들과의 결합을 통해 개인을 식별할 가능성이 있는 속성을 '준식별자(quasi-identifier)'라고 한다. 따라서 데이터 공유에 있어 식별자 제거와 함께 준식별자를 비식별화 할 필요가 있다. 가장 대표적인 프라이버시 모델인 k-익명성은 준식별자에 대하여 동일한 속성 값을 가지는 레코드들의 집합인 동질 클래스(equivalence class)의 크기를 k 이상으로 요구함으로, 특정 레코드를 동질 클래스내의 다른 (k-1)개의 레코드들과 구별할 수 없게 한다 [5,6]. 가령, 그림 1은 원본 테이블과 2-익명화 테이블을 나타내며, 준식별자는 'Gender'와 'Age'에 해당된다. 2-익명화 테이블에는 2개의 동질 클래스(1번~4번 레코드들로 구성된 동질 클래스, 5번~6번 레코드들로 구성된 동질 클래스)가 존재한다. 익명화된 테이블에서 각각의 레코드는 동질 클래스 내의 다른 레코드들과 구분이 되지 않는다. 가령, 1번 레코드는 2-익명화 테이블에서 다른 세 개의 레코드들(RID=2,3,4)과 구별할 수 없으므로, 익명성이 보장된다. k-익명성에서 데이터를 변형하는 대표적인 방법으로는 일반화(generalization) 기법이 사용된다. 준식별자를 일반화 하는 방법에는 전역 일반화 방법과 지역 일반화 방법이 있다 [9].

Original Table

RID	Gender	Age	Disease
1	M	24	Pneumonia
2	M	29	Diabetes
3	F	26	Anemia
4	F	29	Pneumonia
5	M	52	Anemia
6	F	51	Diabetes

2-Anonymized Table

RID	Gender	Age	Disease
1	* (0~1)	20~29	Pneumonia
2	* (0~1)	20~29	Diabetes
3	* (0~1)	20~29	Anemia
4	* (0~1)	20~29	Pneumonia
5	* (0~1)	50~59	Anemia
6	* (0~1)	50~59	Diabetes

Fig. 1. Original table and 2-anonymized table (global generalization)

1.1 Global Generalization

전역 일반화 방법은 원래의 데이터를 전역 일반화 규칙에 따라 상위 일반화 데이터로 대체함으로써 데이터를 익명화하는 방법이다[5,6,9]. 전역 일반화 규칙은 전역 일반화 방법에 적용되는 규칙으로써, 각 속성의 범주 트리(taxonomy tree)를 토대로 만들어진다. 범주 트리는 계층적 트리로서, 상위 단계일수록 높은 일반화 수준을 나타낸다. 예를 들어, 그림 2는 'Age'와 'Gender' 속성에 대한 범주 트리다. 그림 2 에서 보는 것과 같이 'Age'에 대한 범주 트리는 3단계의 일반화 수준을 가지고 있고, 'Gender'에 대한 범주 트리는 2단계의 일반화 수준을 가지고 있다. 그림에서 보듯이 낮은 일반화 수준일수록 원래의 값과 가깝고, 높은 일반화 수준일수록 일반화된 값의 범위가 넓다. 그렇기 때문에 일반화 수준이 낮아지면 데이터의 활용도가 높아지지만 개인정보 노출의 위험도 또한 증가한다. 반대로 일반화 수준이 높아지면 개인정보 노출의 위험도가 낮아지지만 데이터의 활용도 또한 낮아진다. 그래서 전역 일반화 방법에서의 핵심은 적절한 범주 트리의 형태와 그 형태에 따른 적합한 전역 일반화 규칙을 구성하는 것에 있다. 또한, 그림 2의 'Gender'에 해당하는 범주 트리에서 보여 지듯이, 범주형 자료(categorical data)는 수치적 자료(numerical data)로 표현 가능하다.

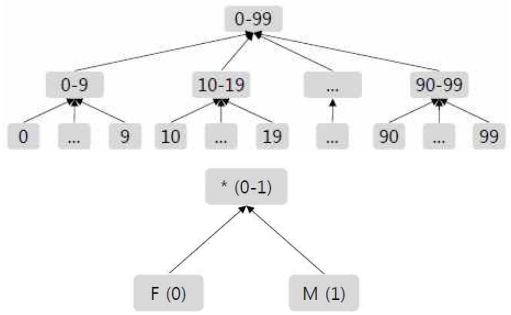


Fig. 2. Example taxonomy trees of Age and Gender

1.2 Local Generalization

지역 일반화 방법은 전역 일반화와 달리 일정한 규칙 없이 데이터들 간의 유사성을 이용하여 일반화하는 방법이다. 그래서 지역 일반화를 적용한 데이터들은 같은 값이라도 각각 다른 형태로 일반화 될 수 있다. 그렇기 때문에 동질 클래스의 크기, 일반화의 범위 등이 각각 다를 수 있다. 예를 들어, 그림 3은 지역 일반화 방법을 이용하여 그림 1의 원본 테이블을 2-익명화한 테이블로서, 3개의 동질 클래스(1번~2번 레코드, 3번~4번 레코드, 5번~6번 레코드로 구성된 동질 클래스)가 존재한다. 2번과 4번 레코드는 원본 테이블에서 'Age' 속성에 대하여 같은 값을 가지고 있지만, 그림에서 보듯이 서로 다른 값으로 일반화 할 수 있다. 또한, 그림 1의 전역 일반화를 이용하여 익명화된 테이블의 경우, 속성 'Gender'의 값은 모두 '*(0~1)'로 일반화 되었다. 그러나 그림 3의 테이블에서는 'Gender' 속성에 대하여 동질 클래스마다 서로 다른 값으로 일반화 된 것을 알 수 있다.

지역 일반화는 군집화 기술을 통해 구현할 수 있다 [9,10]. 군집화를 통한 k-익명화는 데이터 집합을 k개 이상 유사한 레코드로 묶어 군집화함으로써 일반화하는 방법이다 [10]. 일반적으로 지역 일반화는 사용자가 군집의 특성을 어떻게 설정하느냐에 따라 다양한 익명화 수준을 가질 수 있다.

2-Anonymized Table			
RID	Gender	Age	Disease
1	M (1)	24~29	Pneumonia
2	M (1)	24~29	Diabetes
3	F (0)	26~29	Anemia
4	F (0)	26~29	Pneumonia
5	* (0~1)	51~52	Anemia
6	* (0~1)	52~52	Diabetes

Fig. 3. 2-anonymized table with local generalization

2. Problem Definition

원본 테이블 T1, T2에 대하여, k-익명화 테이블을 각각 AT1, AT2라 가정하자. 또한, 준식별자 속성들의 집합을 Q라 가정하자. 본 논문에서는 k-익명화 테이블이 조인 결과에 미치는 영향을 실험적으로 비교 연구한다. 즉, 데이터 사용자는 k-익명화 테이블 AT1, AT2에 대하여 다음과 같은 조인 질의를 요청한다고 가정하자.

```
SELECT   COUNT(*)
FROM     AT1, AT2
WHERE    AT1.q1 = AT2.q1
AND      AT1.q2 = AT2.q2
AND      .......
AND      AT1.qr = AT2.qr
```

이때, $q = \{q_1, q_2, \cdots, q_r\}$ 는 준식별자 속성들의 집합으로서, Q의 부분 집합에 해당한다 (q⊂Q). 위의 질의는 준식별자 속성을 이용하여 k-익명화 테이블 간의 등가 조인을 수행한 후, 조인 결과에 포함되는 레코드의 개수(즉, 조인 카디널리티, join cardinality)를 구하는 질의이다. 본 연구에서 위와 같은 조인 질의를 대상으로 하는 이유는 COUNT() 함수를 포함하는 two-way 조인 질의는 실제 데이터 분석에서 가장 많이 사용되는 질의에 해당되기 때문이다. 또한, 본 논문에서는 일반화 기법(전역 일반화와 지역 일반화)이 조인 결과에 미치는 영향을 실험적으로 비교한다.

III. Computing Join Cardinality with k-Anonymous Tables

일반적으로 테이블 간의 동등 조인은 서로 같은 값을 가지는 레코드에 대해 수행한다. 하지만 익명화된 테이블에서는 일반화된 속성 값으로 인하여 특정 속성의 정확한 값을 얻을 수 없다. 예를 들어, 그림 4와 같이 AT1의 1,2,3번 레코드, AT2의 1,2번 레코드가 각각 동질 클래스를 형성한다고 가정하자. 또한, AT1의 1,2,3번 레코드와 AT2의 1,2번 레코드에 대하여 'Age' 속성을 이용하여 등가 조인을 수행한다고 가정하자. 이 경우 모든 레코드의 'Age' 값은 {20-22}로 일반화 되어있으므로, 단순 방법(naive approach)에 의해 AT1의 모든 레코드들과 AT2의 모든 레코드들 간의 조인이 가능하다. 따라서 두 동질 클래스 간의 조인 카디널리티는 6(=3×2)에 해당한다.

그러나 이러한 단순 방법은 조인 카디널리티를 과대측정(overestimation) 한다는 문제점이 있다. 즉, 동질 클래스에 속하는 레코드들은 준식별자에 대해 서로 다른 값을 가질 수 있으며, 이 경우 실제 조인 카디널리티는 단순 방법에 의해 계산된 조인 카디널리티보다 작게 된다. 그러므로 본 논문에서는 레코드들의 속성 값들은 균등 분포한다는 가정 하에, 다음과 같이 조인 카디널리티를 계산한다. 따라서 위 예에서 속성 'Age' 값은 {20-22}로 일반화 되어있으므로, 가능한 단위 값은 {20}, {21}, {22}이 있다. 그리고 AT1의 1,2,3번 레코드로 구성된 동질 클래스에는 3개의 레코드가 존재하므로, 각각의 단위 값을 가지는 레코드의 수는 평균적으로 1(=3/3)에 해당한다. 이와 유사하게, AT2의 1,2번 레코드로 구성된 동질 클래스의 경우 각각의 단위 값을 가지는 레코드의 수는 평균적으로

0.667(=2/3)에 해당한다. 결국 이 예시에서 조인을 실시할 경우, 같은 단위 값에 대하여 조인이 가능하므로, 조인 카디널리티는 2 ($\approx 1 \times 0.667 \times 3$)로 계산 가능하다. 본 장에서는 위와 같은 방법을 이용하여 일반화된 테이블간의 조인 카디널리티를 계산한다.

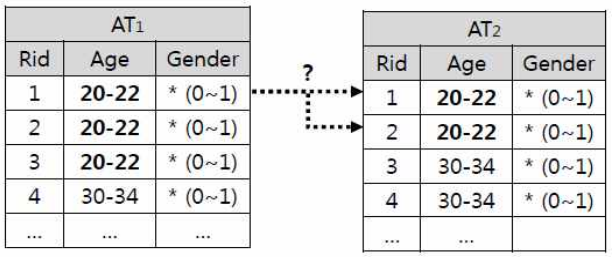


Fig. 4. Join between two k-anonymized tables with global generalization

1. Join between k-anonymous tables with Global Generalization

AT1과 AT2에 각각 다음과 같은 n개와 m개의 동질 클래스가 존재 한다고 가정하자.

$$E_{AT1} = \{e_{\langle 1,1\rangle}, e_{\langle 1,2\rangle}, \cdots, e_{\langle 1,n\rangle}\}$$
$$E_{AT2} = \{e_{\langle 2,1\rangle}, e_{\langle 2,2\rangle}, \cdots, e_{\langle 2,m\rangle}\}$$

특정 속성 $q_s \in q$에 대하여, 익명화 테이블 AT1의 i-번째 동질 클래스 e<1,i>에 속하는 레코드들이 가질 수 있는 최대값과 최소값을 각각 MaxSe<1,i>, MinSe<1,i>라 하자. 이때, 동질 클래스 e<1,i>에 속하는 레코드들이 준식별자 q1, q2, ..., qr에 대해 가질 수 있는 단위 값(unit value)들의 모든 가능한 경우의 수는 다음과 같다.

$$UV_{cnt}(e_{\langle 1,i\rangle}, q) = \prod_{z=1}^{r} (Max^z e_{\langle 1,i\rangle} - Min^z e_{\langle 1,i\rangle} + 1)$$

그러므로 동질 클래스 e<1,i>에 대하여, 준식별자 q1, q2, ..., qr에 대해 가질 수 있는 단위 값 당 평균 레코드 수(Reccnt(·))는 다음과 같다.

$$Rec_{cnt}(e_{\langle 1,i\rangle}, q) = \frac{|e_{\langle 1,i\rangle}|}{UV_{cnt}(e_{\langle 1,i\rangle}, q)}$$

여기서, |e<1,i>|는 동질 클래스 e<1,i>에 속하는 레코드들의 개수를 나타낸다.

동등 조인은 서로 같은 단위 값을 가지는 레코드에 대해 수행 가능하며, 각각의 같은 단위 값 당 조인 카디널리티는 다음과 같이 계산 가능하다.

$$Rec_{cnt}(e_{\langle 1,i\rangle}, q) \times Rec_{cnt}(e_{\langle 2,j\rangle}, q)$$

그러므로 AT1의 i-번째 동질 클래스 e<1,i>와 AT2의 j-번째 동질 클래스 e<2,j> 사이의 조인 카디널리티(JC(·))는 다음과 같다.

$$JC(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}) = Rec_{cnt}(e_{\langle 1,i\rangle}, q) \times Rec_{cnt}(e_{\langle 2,j\rangle}, q) \times Overlap(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle})$$

이때, Overlapcnt(e<1,i>, e<2,j>)는 두 동질 클래스간의 조인 가능한 단위 값들의 개수에 해당하며, 함께 전역 일반화된 테이블 AT1, AT2에 대해 다음과 같다.

$$Overlap_{cnt}(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}) = \begin{cases} UV_{cnt}(e_{\langle 1,i\rangle}, q) \ (= UV_{cnt}(e_{\langle 2,j\rangle}, q)) & (1) \\ 0 & otherwise \end{cases}$$

위에서, (1)의 조건은 다음과 같다.

$$\forall_{1 \le t \le r} \begin{pmatrix} (Min^t e_{\langle 1,i\rangle} = Min^t e_{\langle 2,j\rangle}) \wedge \\ (Max^t e_{\langle 1,i\rangle} = Max^t e_{\langle 2,j\rangle}) \end{pmatrix}$$

즉, 전역 일반화된 테이블의 경우, 일반화된 준속성자 값이 같은 경우 등가 조인이 가능하다.

예제 1. 그림 4와 같이 AT1의 1,2,3번 레코드, AT2의 1,2번 레코드가 각각 동질 클래스 형성하고, 이들을 각각 e<1,1>, e<2,1>라 하자. 또한, q={Age, Gender}라 가정하자. 이때, 동질 클래스 e<1,1>, e<2,1>가 준식별자 'Age', 'Gender'에 대해 가질 수 있는 단위 값들의 모든 가능한 경우는 다음과 같다.

[Age, Gender] = [20, 0], [21, 0], [22, 0],
[20, 1], [21, 1], [22, 1]

$$UV_{cnt}(e_{\langle 1,1\rangle}, q) = UV_{cnt}(e_{\langle 2,1\rangle}, q) = (22-20+1) \times (1-0+1) = 6$$

또한, 준식별자 q={Age, Gender}에 대해 가질 수 있는 단위 값 당 평균 레코드 수는 각각 다음과 같다.

$$Rec_{cnt}(e_{\langle 1,1\rangle}, q) = \frac{3}{6}, \quad Rec_{cnt}(e_{\langle 2,1\rangle}, q) = \frac{2}{6}$$

이를 이용하여, 두 동질 클래스 e<1,1>, e<2,1> 간의 조인 카디널리티는 다음과 같이 구한다.

$$JC(e_{\langle 1,1\rangle}, e_{\langle 2,1\rangle}) = 6 \times \frac{3}{6} \times \frac{2}{6} = 1$$

또한, AT2의 3,4번 레코드로 구성된 동질 클래스를 e<2,2>라 가정하자. 이때, Overlapcnt(e<1,1>, e<2,2>)=0 이므로, e<1,1>, e<2,1> 간의 조인 카디널리티는 다음과 같다.

$$JC(e_{\langle 1,1\rangle}, e_{\langle 2,2\rangle}) = 0 \times \frac{3}{6} \times \frac{2}{6} = 0$$

즉, e<1,1>, e<2,2>에 속하는 레코드들 사이에는 속성 Age 값이 서로 같을 수 없으므로, 등가 조인의 카디널리티는 0에 해당된다.

동질 클래스 간의 등가 조인 카리널리티를 구하는 위 수식을 테이블 AT1, AT2 전체로 일반화하면, 다음과 같이 익명화 테이블 간의 조인 카디널리티를 구할 수 있다.

$$JC(AT_1, AT_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} JC(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle})$$

즉, AT1, AT2 사이의 등가 조인 카리널리티는 모든 동질 클래스 간의 등가 조인 카리널리티의 합계에 해당된다.

2. Join between k-anonymous tables with Local Generalization

지역 일반화된 테이블 간의 조인 카디널리티는 III-1절에서 설명한 전역 일반화 테이블 간의 조인 카디널리티 계산 방법과 유사하게 구한다. 하지만 지역 일반화는 동질 클래스의 일반화 범위가 각각 다르며, 이는 두 동질 클래스간의 조인 가능한 단위 값들의 개수(즉, Overlapcnt(·))에 영향을 미친다. 두 개의 동질 클래스 e<1,i>, e<2,j>가 속성 qs∈q에 대해 중복되는 일반화 값을 형성되는 경우는 다음과 같다 (그림 5의 (a), (b), (c)).

(a) 범위가 서로 같은 경우
(b) 하나의 범위가 다른 범위에 포함되는 경우
(c) 하나의 범위가 다른 범위와 부분적인 공통범위가 있는 경우

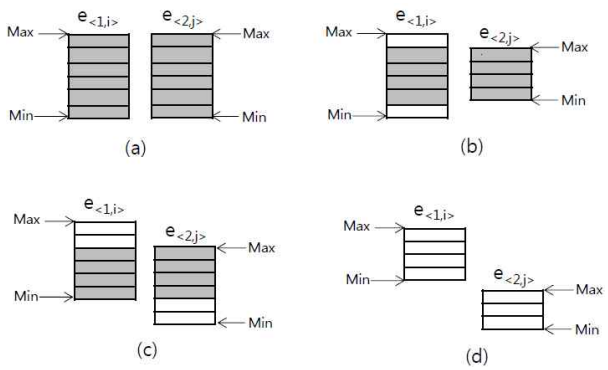


Fig. 5. Possible cases between e<1,i> and e<2,j> with respect to the attribute qs (local generalization): overlap cases (a,b,c) and non-overlap case (d)

동질 클래스 e<1,i>, e<2,j>가 속성 qs∈q에 대해 중복되는 일반화 값을 형성하는 경우 (그림 5의 (a,b,c)), 중복 범위의 최대값과 최소값은 다음과 같다.

$$Min(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}, q_s) = MAX(Min^s e_{\langle 1,i\rangle}, Min^s e_{\langle 2,j\rangle})$$
$$Max(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}, q_s) = MIN(Max^s e_{\langle 1,i\rangle}, Max^s e_{\langle 2,j\rangle})$$

이를 이용하여, 두 동질 클래스간의 조인 가능한 단위 값들의 개수 Overlapcnt(e<1,i>, e<2,j>)는 다음과 같다.

$$Overlap_{cnt}(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}) = \begin{cases} \prod_{z=1}^{r} (Max(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}, q_z) - Min(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}, q_z) + 1) & (1) \\ 0 & otherwise \end{cases}$$

위에서, (1)의 조건은 앞에서 언급한 일반화 값을 형성 조건 (a), (b), (c)에 해당하며, 다음과 같이 표현가능하다.

$$\forall_{1 \le t \le r} \left(\begin{array}{c} \neg((Min^t e_{\langle 1,i\rangle} > Max^t e_{\langle 2,j\rangle}) \vee \\ (Max^t e_{\langle 1,i\rangle} < Min^t e_{\langle 2,j\rangle})) \end{array} \right)$$

즉, 위의 조건은 그림 5-(d)를 제외한 모든 경우를 의미한다.

III-1절과 유사하게, 지역 일반화된 테이블 AT1의 i-번째 동질 클래스 e<1,i>와 AT2의 j-번째 동질 클래스 e<2,j> 사이의 조인 카디널리티는 다음과 같이 계산 가능하다.

$$JC(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}) = Rec_{cnt}(e_{\langle 1,i\rangle}, q) \times Rec_{cnt}(e_{\langle 2,j\rangle}, q) \times Overlap(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle})$$

여기서 단위 값 당 평균 레코드 수 Reccnt(·)는 III-1절과 동일한 방법으로 구한다.

예제 2. 그림 6와 같이 AT1의 1,2,3번 레코드, AT2의 1,2번 레코드가 각각 동질 클래스 형성하고, 이들을 각각 e<1,1>, e<2,1>라 하자. 또한, q={Age, Gender}라 가정하자. 이때, 동질 클래스 e<1,1>가 준식별자 'Age', 'Gender'에 대해 가질 수 있는 단위 값들의 모든 가능한 경우는 다음과 같다.

[Age, Gender] = [20, 0], [21, 0], [22, 0],
[20, 1], [21, 1], [22, 1]

또한, 동질 클래스 e<2,1>가 준식별자 'Age', 'Gender'에 대해 가질 수 있는 단위 값들의 모든 가능한 경우는 다음과 같다.

[Age, Gender] = [21, 0], [22, 0], [23, 0],
[21, 1], [22, 1], [23, 1]

그러므로 e<1,1>, e<2,1>가 가질 수 있는 단위 값이 주어졌을 경우, 실제 조인 가능한 경우의 수는 다음과 같다.

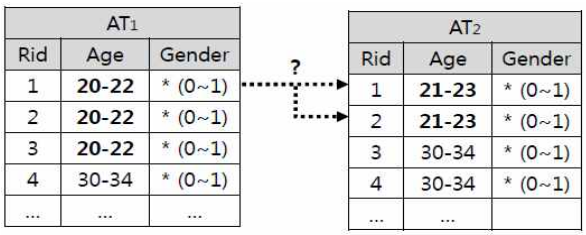
[Age, Gender] = [21, 0], [22, 0], [21, 1], [22, 1]

$$Overlap_{ant}(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle}) = (22-21+1)\times(1-0+1) = 4$$

이를 이용하여, 두 동질 클래스 e<1,1>, e<2,1> 간의 조인 카디널리티는 다음과 같이 구한다.

$$JC(e_{\langle 1,1\rangle}, e_{\langle 2,1\rangle}) = 4 \times \frac{3}{6} \times \frac{2}{6} = 0.667$$

지역 일반화로 인하여 'Age'속성의 일반화 값이 부분적으로 포함하는 관계이므로, 예제 1과 비교하여 두 동질 클래스 간의 조인 카디널리티가 변화한 것을 알 수 있다.



AT1

Rid	Age	Gender
1	20-22	* (0~1)
2	20-22	* (0~1)
3	20-22	* (0~1)
4	30-34	* (0~1)
...	...	...

?

AT2

Rid	Age	Gender
1	21-23	* (0~1)
2	21-23	* (0~1)
3	30-34	* (0~1)
4	30-34	* (0~1)
...	...	

Fig. 6. Join between two k-anonymized tables with local generalization

마지막으로, 테이블 AT1, AT2 간의 조인 카디널리티는 다음과 같다.

$$JC(AT_1, AT_2) = \sum_{i=1}^{n}\sum_{j=1}^{m} JC(e_{\langle 1,i\rangle}, e_{\langle 2,j\rangle})$$

즉, III-1절과 동일하게, 테이블 AT1, AT2 간의 조인 카디널리티는 모든 동질 클래스 간의 등가 조인 카리널리티의 합계에 해당한다.

IV. Experiment

본 절에서는 논문에서 제안한 기법의 성능 평가를 보건의료 데이터[11]를 사용하여 수행한다. 본 실험에서는 보건의료 데이터로부터 'Age', 'Sex', 'Location', 'Surgery', 'Length', 'Disease' 속성을 추출하여 원본 테이블 T1, T2를 각각 생성하였다. 이때, 'Disease' 속성은 민감한 속성(sensitive attribute)에 해당하며, 그 이외의 속성들은 준식별자에 해당한다. 또한 데이터의 익명화 방법으로는 [6](전역 일반화)과 [7](지역 일반화)의 알고리즘을 사용하였다.

본 실험에서는 II-2절의 사용자 질의가 주어졌을 때, 다음과 같은 방법을 이용하여 조인 카디널리티를 구한다.

전역 일반화된 테이블에 대한 조인 (GG, Global Generalization) - III-1절에 설명한 방법으로서, 전역 일반화된 테이블에 대하여 조인 수행

지역 일반화된 테이블에 대한 조인 (LG, Local Generaliztion) - III-2절에 설명한 방법으로서, 지역 일반화된 테이블에 대하여 조인 수행

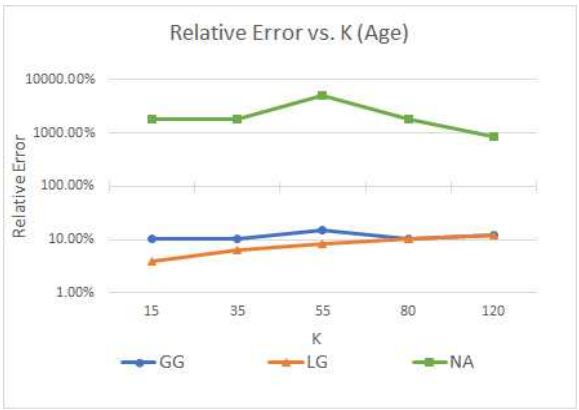
단순 방법 (NA, Naive Approach) - III절에서 언급한 단순 방법으로서, 전역 일반화된 테이블에서 같은 값으로 일반화된 레코드 사이에 조인 수행

본 실험에서의 성능 비교를 위하여 다음과 같이 상대오차 (relative error)를 측정하였다.

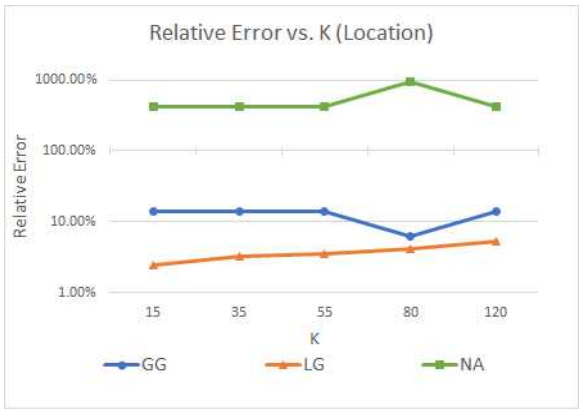
$$\frac{|\text{원본 데이터 조인 카디널리티} - \text{일반화 데이터 조인 카디널리티}|}{\text{원본 데이터 조인 카디널리티}} \times 100$$

1. Experimental Results

그림 7은 조인 속성이 1개인 경우, k 값의 변화에 따른 상대오차를 측정한 실험이다. 이 실험에서 사용한 테이블 T1, T2는 각각 100,000개의 레코드들로 구성되어있으며, 조인 속성으로 'Age'와 'Location'을 사용하였다. 그리고 k 값은 실험에서 15,



(a) Join attribute = 'Age'



(b) Join attribute = 'Location'

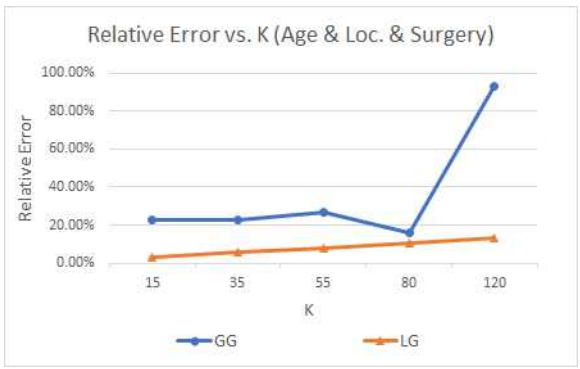
Fig. 7. Relative error on varying k
(The Number of join attributes = 1)

35, 55, 80, 120로 변화 시키면서, 상대오차를 측정하였다.

그림 7의 (a)와 (b)에서 보듯이 모든 k 값에 대하여, LG 기법이 GG 기법보다 낮은 상대오차를 보이고 있다. 이는, 데이터 유용성 측면에서, 일반적으로 지역 일반화 기법이 전역 일반화 기법보다 우수하다고 알려져 있고[9,10], 이러한 데이터 유용성에서의 차이는 조인 결과에 영향을 미치기 때문이다. 또한 두 방법 모두 k 값이 증가함에 따라 상대오차가 증가하는 추세를 보였다. 이는 k 값이 증가함에 따라 데이터 유용성이 감소하고, 이로 인하여 조인 결과에 영향을 미치기 때문이다. 또한, 두 기법 간의 성능 차이는 대체적으로 k 값이 증가함에 따라, 작아지는 것을 알 수 있다. 꾸준히 증가함을 보이는 LG 기법과는 달리 GG 기법은 k 값이 증가에 따라 상대오차가 감소했다가 다시 증가하는 현상을 그림 7-(b)에서 보였다.

그림 7에서 보듯이, 단순 방법인 NA는 조인 결과의 과대측정으로 인하여, 모든 k 값에 대하여 매우 높은 상대오차를 보이고 있다. 이는 익명화된 테이블에 대한 조인 질의는 본 논문에서 제안한 방법과 같이 동질 클래스에 속하는 레코드들을 각각의 단위 값으로 분배한 후, 조인을 수행할 필요가 있음을 실험적으로 증명해 주고 있다.

(a) Join attribute = 'Age', 'Location'



(b) Join attribute = 'Age', 'Location', 'Surgery'

Fig. 8. Relative error on varying k
(The Number of join attributes = 2, 3)

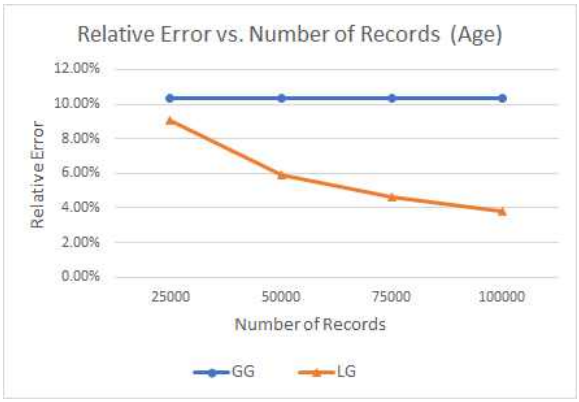


Fig. 9. Relative error on varying the number of records in T1 and T2

그림 8는 2개 이상의 속성으로 조인을 실시한 결과이다. 그림 8-(a)에서는 'Age', 'Location' 속성을 사용하였고, 8-(b)에서는 'Age', 'Location', 'Surgery' 속성을 사용하여 조인을 실시했다. 이 실험에서 사용한 테이블 T1, T2는 각각 100,000개의 레코드들로 구성되어있으며, 실험에서 k 값을 15, 35, 55, 80, 120로 변화 시키면서, 상대오차를 측정하였다. 단순 방법인 NA의 경우 이전 실험에서 매우 높은 상대오차를 보였으므로, 본 실험부터는 GG 기법과 LG 기법만을 이용하여 사용자 질의를 수행하였다. 그림 8에서 보듯이 두 실험 결과 모두 LG 기법이 GG 기법보다 낮은 상대오차를 보이고 있다. 또한, 두 방법 모두 k 값이 증가함에 따라 상대오차가 증가하는 추세를 보였다. 이는 이전 실험 결과와 같이 k 값이 증가함에 따라 데이터 유용성이 감소하고, 이로 인하여 조인 결과에 영향을 미치기 때문이다.

마지막으로, 그림 9는 테이블 T1, T2에 속하는 레코드의 수를 25,000개에서 100,000개까지 25,000개씩 증가시켰을 때, 상대오차를 측정한 실험 결과이다. 이 실험에서 k 값은 15이고, 조인 속성은 'Age'를 사용하였다. 그림 9에서 보여지듯이 모든 데이터 크기에서 LG 기법이 GG 기법보다 낮은 상대오차를 보이고 있다. LG 기법은 데이터 크기가 증가함에 따라 상대오차가 꾸준하게 감소하는 추세를 보인다. 즉, 테이블에 속하는 레코드 수가 증가 할수록 지역 일반화는 익명화되기 이전의 원본 테이블에 대해 조인 질의를 수행한 결과와 유사해 진다. 반면에 전역 일반화는 레코드 수가 증가하여도 상대오차가 일정하다.

본 절에서의 실험 결과는 다음과 같이 요약 할 수 있다. 익명화된 테이블에 대한 조인 질의는 본 논문의 III 절에서 제안한 방법 (즉, 동질 클래스에 속하는 레코드들을 각각의 단위 값으로 분배한 후, 조인을 수행하는 기법)을 통해, 효율적으로 수행할 수 있다. 또한, 지역 일반화된 테이블에 대한 조인 결과가 전역 일반화된 테이블에 대한 조인 결과보다 우수하다. 이는 데이터 유용성 측면에서, 지역 일반화 기법이 전역 일반화 기법보다 우수하고, 데이터 유용성의 차이는 조인 결과에 직접적으로 영향을 미친다는 것을 나타낸다.

V. Conclusion and Future Work

본 논문에서는 k-익명화 테이블에 대한 조인 질의 결과의 카디널리티를 구하기 위한 방법을 제안하였다. 본 논문에서 제시한 방법은 동질 클래스에 속하는 레코드들을 각각의 단위 값으로 분배한 후, 조인을 수행하는 방법이다. 또한, 실 데이터를 이용한 실험을 통하여, 본 논문에서 제안한 기법이 단순 방법보다 우수함을 알 수 있었다. 특히, 서로 다른 일반화 기법간의 비교 실험에서는 지역 일반화된 익명화 테이블에 대한 조인 결과가 전역 일반화된 익명화 테이블에 대한 조인 결과 보다 우수함을 알 수 있었다. 이는 익명화된 테이블에 대한 조인 질의를 필요로 하는 응용 프로그램의 경우, 지역 일반화된 익명화 테이블이 전역 일반화된 익명화 테이블보다 더 유용하다는 것을 암시한다.

향후 연구 계획으로는 조인 질의의 종류를 확장할 계획이다. 또한, two-way 조인을 확장하여, n-way 조인을 수행하기 위한 방법에 관한 연구를 진행할 예정이다.

References

[1] A. Narayanan and V. Shmatikov, “Robust De-anonymization of Large Sparse Datasets”, In Proceedings of the 2008 IEEE Symposium on Security and Privacy Page, 2008.

[2] J. Kim, K.Jung, H. Lee, S. Kim, J.W. Kim and Y.D. Chung, “Models for Privacy-preserving Data Publishing : A Survey”, Journal of KIISE, Vol. 44, No. 2, pp. 195-207, 2017.

[3] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, “Privacy-preserving data publishing: A survey of recent developments”, ACM Computing Surveys, 42(4), June 2010.

[4] N. Mohammed, B.C.M. Fung, P.C.K. Hung, and C.K. Lee, “Centralized and distributed anonymization for high-dimensional healthcare data”, ACM Transactions on Knowledge Discovery from Data, 4(4), October 2010.

[5] L. Sweeney, “k-anonymity: A model for protecting privacy”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5), 557-570, 2002.

[6] K. LeFevre, D.J. DeWitt and R. Ramakrishnan, “Incognito: Efficient full domain k-anonymity”, In Proceedings of the ACM SIGMOD International Conference on Management of Data, 2005.

[7] A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity”, ACM Transactions on Knowledge Discovery from Data, 1(1), 2007.

[8] N. Li, T. Li and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity”, In Proceedings of the International Conference on Data Engineering, 2007.

[9] S. Kim, H. Lee, Y.D. Chung, “Privacy-preserving data cub for electronic medical records: An experimental evaluation”, International Journal of medical Informatics, 2017

[10] J. Byun, A. Kamra, E. Bertino, N. Li, “Efficient k-Anonymization Using Clustering Technique”, DASFAA 2007: Advances in Databases: Concepts, Systems and Applications pp 188-200, 2007

[11] Health Insurance Review and Assessment Service in Korea. http://opendata.hira.or.kr (2012).

Authors

Dae Ho Kim is a graduate student in computer science at Sangmyung University. Kim received his BS degree in Media Software from Sangmyung University in 2017.His interests are Big Data, Data Mining and Distributed Computing.

Jong Wook Kim is an assistant professor of Computer Science at Sangmyung University. His research interests include Web data mining, information retrieval, and database systems. Kim has a PhD from the School of Computing, Informatics, and Decision Systems Engineering at Arizona State University. He was a software engineer in Teradata Corporation. He is a member of the IEEE and the ACM.