상관도를 이용한 국내 의료기관용 개인정보 비식별화 방안에 관한 연구

Considering on De-Identification Method of Personal Information for National Medical Institute by using correlation

여광수*, 김철중*, 이재현**,김순석*

(Kwang Soo Yeo*, Chul Jung Kim*, Jae Hyun Lee**, Soon Seok Kim*)

요 약

의료기관의 개인정보들을 보호하기 위한 가이드라인은 미국, 영국 등 각 나라에서도 이미 진행되어온 상태이고 또한 HIPPA와 같이 여러 곳에서 발표 되고 있다. 하지만 국내의 경우, 국내 의료기관에 특화된 가이드라인에 대해서는 명확히 제시되지 않고 있는 실정이다. 본 논문은 지난 2015년 미래창조과학부에서 발표된 빅데이터 비식별화 기술 활용 안내서를 기반으로 국외인 영국의 ICO, 미국의 IHE, NIST, HIPPA에서 발표한 의료기관의 개인정보보호 비식별화 기술에 관한 가이드들을 고찰하여 국내 의료기관에서 활용할 수 있는 기술적인 방안과 상관도를 제시하였다. 여기서 상관도란 앞서 미국의 3개 기관에서 제시하고 있는 기술들에 대해 공통적으로 제시하고 있는 정도를 5점 척도로 나타낸 것을 의미한다. 즉, 5점에 가까울수록 여러 기관들에서 제시된 기술을 활용할 것을 많이 권고한다는 의미이다. 본 논문을 통하여 기초 자료로서 국내 의료기관에서 개인의료정보 비식별화에 더 많은 발전과 활용이 되기를 바란다.

■ 중심어 : 개인건강정보 ; 비식별화 ; 프라이버시 보호 ; 민감정보 ; 의료정보

Abstract

Guidelines for protecting personal information are already in progress in USA, UK and other countries and announced many guideline like HIPPA. However In Our national environment, we does not have specialized guideline in national medical industries. This thesis suggest De-indentification method in South Korea by referring 'bigdata De-identification Guideline by Ministry of Science, ICT and Future Planning (2015)', ICO in U. K and IHE, NIST, HIPPA in U. S. A. We suggest also correlation between Guidelines. Corelation means common techniques in three guidelines (IHE, NIST, HIPPA in U. S. A). As Point becomes closer five points, We recommend that technique to national medical institute for De-Identification. We hope this thesis makes the best use of personal information's development in National medical institute.

■ keyword : Personal Health Information ; De-Identification ; Privacy Protection ; Sensitive Information ; Medical Information

Ⅰ. 서 론

본 논문에서는 최근 들어 더욱 중요성이 높아지고 있는 의료기관의 개인정보보호 기술들에 대해 다루고자 한다. 국내 의료법 상 의료기관에서 환자의 의료정보를 1차 이용하는 것뿐만 아니라 특히 개인정보보호법과 의료법의 발효로 의료기관 내에서 연구 또는 교육 등을 목적으로 환자 개인의료정보를 2차 이용할 경우 비식별 화가 반드시 법적으로 요구되는 것이 현실이다. 이러한 개인정보들을 보호하기 위한 가이드라인은 미국, 영국 등 각 나라에서도 이미 진행이 되어져 있는 상태이고 여러 곳에서 발표 되고 있다. 그러나 국내의 경우, 현재까지 국내 의료기관에 특화된 가이드라인에 대해서는 명확히 제시되지 않고 있는 실정이다.

본 논문에서는 지난 2015년 미래창조과학부에서 발표된 빅데이터 비식별화 기술 활용 안내서[1]를 기반으로 영국의 ICO[2], 미국의 IHE[3], NIST[4], HIPPA[5]에서 발표한 의료기관의 개인정보보호 비식별화 기술들에 관한 가이드들을 검토하여 국내 의료기관에서 활용할 수 있는 기술적인 방안과 연관도를 제시하고자 한다.

* 정회원, 한라대학교 컴퓨터공학과

접수일자 : 2016년 09월 19일
수정일자 : 2016년 12월 09일
게재확정일 : 2016년 12월 26일
교신저자 : 김순석 e-mail : sskim@halla.ac.kr

(표 1) 국내 의료기관을 위한 국내외 개인정보 비식별화 방안

비식별화 기법		국내 [1]	영국 [2]	미국			의료분야 적용가능 기술	상관도 (5점 척도)
				NIST[4]	IHE[3]	HIPAA[5]		
가명처리 Pesudonymization	휴리스틱 익명화 Huristic Pseudonymization	○	X	X	X	X	X	0
	K-익명화 K-anonymity	○	X	○	X	X	○	2
	암호화 Encrytion	○	X	X	△	X	△	1
	교환방법 Swapping	○	X	○	X	X	○	2
총계처리 Aggregation	총계처리 Aggregation	○	○	X	○	X	○	2
	부분집계 Micro Aggregation	○	X	X	X	X	X	0
	라운딩 Rounding	○	○	X	○	X	○	2
	데이터 재배열 Rearrangement	○	X	X	○	X	○	2
데이터 값 삭제 Data Reduction	속성값 삭제 Reducing Variables	○	X	○	○	○	○	6
	속성값 부분 삭제 Reducing Partial Variables	○	○	○	○	○	○	6
	데이터 행 삭제 Reducing Records	○	X	X	X	X	X	0
	식별자 제거를 통한 단순 익명화 Trivial Anonymization	○	X	X	X	X	X	0
범주화 Data Suppression	범주화 Data Suppression	○	○	○	△	X	○	3
	랜덤 올림 방식 Random Rounding	X	X	X	X	X	X	0
	범위 방법 Data Range	○	X	X	○	X	○	2
	제어 올림 Controlled Rounding	○	X	X	X	X	X	0
데이터 마스킹 Data Masking	임의 잡음 추가 Adding Random Noise	○	X	X	X	X	X	0
	공백 Blank & 대체 Impute	○	○	X	○	X	○	2

본 논문의 2장에서는 의료기관을 위한 국내외의 비식별화 관련 동향들을 살펴보고, 3장에서는 국내 의료기관을 위한 개인정보 비식별화 방안을 제안한 후, 4장을 끝으로 결론을 맺고자 한다.

II. 관련연구

개인정보 비식별화와 관련하여 국내에서 제시하고 있는 대표적인 방법은 지난 2015년 미래창조과학부에서 발표한 가이드라인이 대표적이며 그 외 영국, 미국 등 각 나라에서도 아래 (표 1))에서와 같이 기술적인 가이드라인을 제시하고 있다.

(표 1)은 이들 각 나라의 기술적인 가이드라인들을 비교한 것이며, 국내와 국외(영국, 미국)의 가이드라인에서 제시하고 있는 기법들을 모두 (표 1)에서 제시하였다. 국내의 경우는 빅데이터 비식별화 기술 활용 안내서[1]를 토대로 하였으며 영국[2]의 경우는 ICO[2]의 내용을 참고하였다.

1. 국내 비식별화 동향

빅데이터 비식별화 기술 활용 안내서[1]을 기반으로 한 국내 비식별화 동향을 살펴보면 제시된 큰 범주 5가지 중 3가지의 기법을 적용 시킨다고 할 수 있다. 이들 중 가명처리(Pseudonymization), 총계처리(Aggregation), 데이터 값 삭제(Data Reduction), 데이터 마스킹 (Data Masking)의 기법이 위의 4가지에 해당되며, 나머지 기법인 범주화(Data Suppression) 기법 4가지 중 랜덤 올림 방식 (Random Rounding)을 제외한 모든 기법이 적용되는 것을 알 수 있다. 그러나 이러한 기법들은 의료기관이 아닌 일반적인 환경에 적합한 기법들을 제시한 것으로 의료법 등 국내의 법과 의료 환경에 적합한 최적화된 비식별화 기법은 아직 제시되고 있지 않은 실정이다.

2. 해외 비식별화 동향

해외에서의 비식별화 동향은 영국의 ICO[2], 미국의 IHE[3], NIST[4], HIPPA[5]를 기반으로 조사하였다. 우선 영국 ICO[2]의 경우 (표 1)에 나오는 큰 범주 5가지를 모두 만족 시킨다고 볼 수 없으며 세부 기법들 중 1개 정도 씩만 만족하고 다른 기법들은 적용하고 있지 않다. 미국의 경우 또한 (표 1)에 나오는 큰 범주 5가지를 모두 만족 시킨다고 볼 수 없다.

미국의 가이드라인 3가지를 분석한 결과, 각 가이드라인에 따라 다르기는 하지만 각 기법들을 모두 만족 시킨다고 볼 수 없으며 그중 공통적으로 적용하고 있는 기법은 데이터 값 삭제 (Data Reduction) 기법이다. 데이터 값 삭제 기법 중에서도 속성 값 삭제 (Reducing Variables) 기법과 속성 값 부분 삭제 (Reducing Partial Variables) 기법의 경우는 모든 가이드라인을 만족하는 기법이다. 그러나 이 두 가지 기법의 경우는 비식별화의 성능은 우수하나 재식별이 쉽지 않는 즉, 유용성이 떨어지는 문제를 가지기 때문에 상호 배타적이라 볼 수 있다. 그래서 미국의 HIPPA 가이드라인[5]에서는 '전문가 활용 기법 (Expert METHOD)'을 제시하고 있는데 우리는 이 기법을 이용하여 신뢰도 있는 전문가가 각 환경에 맞는 규칙을 정하고 그 규칙에 맞는 비식별화를 진행하는 기법을 추천한다. 이 경우 보안성과 유용성이라는 상호 배타적인 관계를 잘 조율하여 융통성 있게 비식별화를 할 수 있다는 장점을 가진다.

3. 가이드라인 적용 여부 및 상관도

(표 1)에서 제시하고 있는 바와 같이, 각 가이드라인 적용 여부에 따라 O또는 X로 표기하였으며 일부 부분 적용의 경우 △로 표기하였다. 또한 이들을 토대로 국내 의료기관에서 적용, 활용할 수 있는 비식별화 기술의 가부를 표기하였고 상관도에 따라 5점 척도로 그 관련성을 표기하였다. 여기서 상관도란 앞서 미국의 3개 기관에서 제시하고 있는 기술들에 대해 공통적으로 제시하고 있는 정도를 5점 척도로 나타낸 것으로 O의 개수에 따라 한 개를 만족하면 2점, 2개를 만족하면 4점, 3개를 만족하면 6점을 부여하였다. 또한 △의 경우는 1점을 부여하였다. 즉, 6점에 가까울수록 여러 기관들에서 제시된 기술을 활용할 것을 많이 권고하고 있다는 의미이다.

4. 국내외 비식별화 기법 소개

앞서 살펴본 바와 같이, 현재 국내외 비식별화에 관한 기법들은 가명처리(Pesudonymiazation), 총계처리(Aggregation), 데이터 값 삭제(Data Reduction), 범주화(Data Suppression), 데이터 마스킹(Data Masking)과 같이 5가지로 나눌 수 있으며 각각 세부적인 기법들이 존재한다.

본 절에서는 (표 1)에서 보이는 5가지 기법들에 대해 설명하고. 상관도에 따라 2점 이상을 받은 기법들을 설명하고자 한다. 특별히 3점 이상을 받은 기법에 대해서는 다음 장에서 설명 하고자 한다.

가. 가명처리 (Pesudonymization)[1]

가명처리는 개인정보를 타인이 보지 못하도록 다른 무언가로 바꾸는 작업을 의미하는데 휴리스틱 익명화(Huristic Pseudonymization), K-익명화(K-anonymity), 암호화(Encryption), 교환(Swapping) 방법이 있다. 가명처리의 장점으로는 보안에 안전하다는 장점을 가지지만 상호 배타적인 관

계를 갖게 되는데, 다시 말해 유용성이 낮아지게 된다. 본 절에서는 상기 4가지의 세부 기법들 중 국내와 미국의 가이드라인에 포함되어 있고 의료분야에 적용 가능한 기법인 K-익명화와 암호화 방법, 그리고 교환 방법에 대해서 다루고자 한다.

(1) K-익명화(K-anonymity)[1]

공개된 데이터에 대한 연결공격(linkage attack)을 방어하기 위해 제안된 프라이버시 보호 모델이다. K-익명화의 정의로는 주어진 데이터 집합에서 준 식별자 속성 값들이 동일한 레코드가 적어도 K개 존재해야 하는 것이다. (그림 1)은 아무런 비식별화가 되지 않은 공개된 의료 데이터이다. (그림 2)는 K=4인 K-익명화 기법으로 비식별화가 되어있는 그림이다. (그림 2)를 보게 되면 K=4가 적용된 모습을 확인 할 수 있다. 데이터 집합의 일부를 수정하여, 모든 레코드가 자기 자신과 동일한 K-1개 이상의 레코드를 가지고 있는 모습을 볼 수 있고 (그림 2)에서 보게 되면 1-4, 5-8, 9-12의 레코드는 서로 구별되지 않음을 확인 할 수 있다. 따라서 익명화된 데이터 집합에서는 공격자가 정확히 어떠한 레코드가 자신이 원하는 공격 레코드인지 확인하기 어렵게 할 수 있어 프라이버시 보호가 될 수 있다. 여기서, 같은 준 식별자 속성 값들로 익명화된 레코드들의 모임을 '동일 준 식별자 속성 값 집합(equivalent class, 이하 동질 집합)이라고 한다. K-익명성의 문제는 역시 상호 배타적인 관계라는 점이다. 즉, K의 값이 올라갈수록 비식별화는 잘 되지만 그 반면에 유용성이 떨어진다고 볼 수 있다.

	준식별자			민감한 정보
	지역 코드	연령	성별	질병
1	11111	28	남	고혈압
2	11111	21	남	고혈압
3	11111	29	여	위암
4	11111	23	남	위암
5	12345	51	여	간압
6	12345	54	남	고혈압
7	12345	55	여	위암
8	12345	56	남	위암
9	11111	47	남	간암
10	11111	46	여	간암
11	11111	44	남	간암
12	11111	45	여	간암

그림1. 공개 의료 데이터

	준식별자			민감한 정보
	지역 코드	연령	성별	질병
1	111**	<40	*	고혈압
2	111**	<40	*	고혈압
3	111**	<40	*	위암
4	111**	<40	*	위암
5	1234*	>50	*	간압
6	1234*	>50	*	고혈압
7	1234*	>50	*	위암
8	1234*	>50	*	위암
9	111**	4*	*	간암
10	111**	4*	*	간암
11	111**	4*	*	간암
12	111**	4*	*	간암

그림2. 4-익명성 모델에 의해 익명화된 의료데이터

(2) 암호화(Encryption) 방법[1]

암호화 방법은 정보의 가공에 있어 일정 규칙의 알고리즘을 적용하여 암호화함으로써 개인정보를 대체하는 방법이다. 통상적으로 다시 유용하게 사용하기 위해 복호화가 가능하도록 암/복호화 비밀키 값을 가지고 있어 만약 비밀키가 노출될 경우 비식별화의 위협이 존재한다. 따라서 비밀키에 대한 보안 방안도 함께 필요하다고 할 수 있다.

(3) 교환 (Swapping) 방법[1]

교환 방법은 추출된 표본 레코드에 대하여 이루어진다. 미리 정해진 변수(항목)들의 집합에 대하여 데이터베이스의 레코드와 연계하여 교환하는데 이는 안전하다고 볼 수 없다. 즉, 미리 정해진 변수(항목)들의 집합에 대한 데이터베이스의 레코드가 보안에 취약한 경우 역시 비식별화에 대한 위협이 존재한다.

나. 총계처리(Aggregation)[1]

총계처리란 개인 식별이 가능한 데이터에 대하여 통계값(전체 혹은 부분)을 적용하여 특정 개인을 식별 할 수 없도록 하는 것이다. 총계처리의 대상 정보로는 개인과 직접 관련된 날짜정보, 기타 교유 특징(수입, 신체정보, 지료기록, 병력정보, 의료기

록 등)에 대해서 총계처리가 가능하다. 총계처리의 장점으로는 민감한 정보에 대해서 비식별화가 가능하고 다양한 통계분석(전체 혹은 부분)용 데이터 셋 작성에 유리하다는 장점을 가진다. 이 반면 단점은 집계 처리된 데이터를 기준으로 정밀한 분석이 어려우며 집계 수량이 적을 경우에는 데이터 결합 과정에서 개인정보의 추출 또는 예측이 가능하다는 단점을 가지고 있다. 이 장에서는 4가지의 세부 기법 중 국내와 미국의 가이드라인에 포함되어 있고 의료분야에 적용 가능한 기법 3가지인 총계처리 (Aggregation) 기본, 라운딩(Rounding), 데이터 재배열(Rearrangement)에 대해서 다루고자 한다.

(1) 총계처리(Aggregation)의 기본 방식

수집된 정보에 민감한 개인정보가 있을 경우 데이터 집단 또는 부분으로 집계(총합, 평균 등) 처리를 하여 민감성을 낮추는 방식으로 기본적인 총계처리에 해당한다. 위에서 언급 했듯이 집계 수량이 적을 경우에는 데이터 결합 과정에서 개인정보의 추출 또는 예측이 가능하다는 단점을 가지고 있다.

(2) 라운딩(Rounding) 기법[1]

이 방식은 집계 처리된 값에 대하여 라운딩(올림, 내림, 사사오입) 기준을 적용하여 최종 집계 처리방식이다. 일반적으로 총계 처리 기본방식에서 많이 쓰이는 값으로 세세한 정보보다는 전체 통계정보가 필요한 경우 많이 사용한다. 예를 들어, 23, 41, 57, 26, 33 등 세세한 나이의 속성 값을 20, 30, 40 등의 각 대표 연령대로 표기하거나, 3,576,000원, 4,210,000원 등의 소득 표기를 십만 원 혹은 백만 원 단위 이하를 절삭하여 3백만 원, 4백만 원 등으로 집계 처리하는 방식이다. 범주화의 랜덤 올림 방법(random rounding)과도 방식이 유사하여 같은 의미로 사용하기도 한다.

(3) 데이터 재배열 (Rearrangement)[1]

이 방식은 기존 정보 값은 유지하면서 개인 정보와 연관이 되지 않도록 해당 데이터를 재배열 즉, 개인의 정보가 타인의 정보와 뒤섞임으로써 전체 정보의 손상 없이 개인의 민감 정보가 해당 개인과 연결되지 않도록 하는 방법이다.

다. 데이터 값 삭제[5]

데이터 값 삭제는 의료정보에서 개인정보 식별이 가능한 특정 데이터 값을 삭제하는 비식별화 기법이다. 일반적으로 데이터 값을 삭제하는 대상은 개인식별정보와 고유식별정보(이름, 전화번호, 주소, 생년월일, 사진, 주민등록번호 등)와 같은 직접 식별자 또는 민감한 데이터이다. 데이터 값 삭제를 통한 비식별화는 식별 가능한 정보에 대한 완전한 삭제 처리가 가능하기 때문에 공격자가 데이터를 예측, 추론 등을 하기 어렵다는 장점이 있다. 그러나 데이터의 완전 삭제로 인해 개인의료 민감 정보의 2차이용 시 데이터 신뢰성, 결과 유용성 그리고 분석 다양성의 저하가 생길 수 있다는 단점이 존재한다. 데이터 값 삭제의 세부 기술은 속성 값 삭제, 속성 값 부분 삭제, 데이터 행 삭제, 식별자 제거를 통한 단순 익명화가 있다.

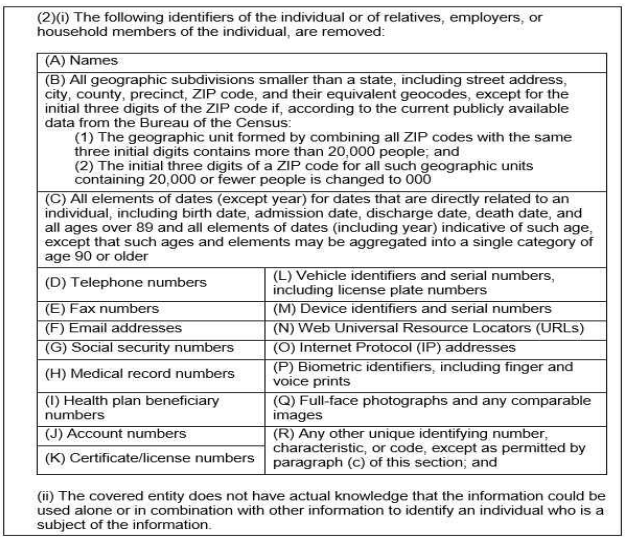
III. 국내의료기관을 위한 개인정보 비식별화 방안

앞서 (표 1)에서 살펴본 바와 같이, 본 연구 내용에서 높은 점수를 받은 항목은 다음과 같이 총 3가지 항목이다. 속성 값 삭제(Reducing Variables), 속성 값 부분 삭제(Reducing Partial Variables), 그리고 범주화 기본(Data Suppression)이 이에 해당되는데 이 세 가지 항목이 상관도에 따라 점수를 각각 3~5점을 받은 고득점의 기법들로 우리나라의 국내 의료기관을 위한 개인정보 비식별화에 적합한 항목들로 조사되었다.

이 절에서는 높은 득점을 받고 국내의 의료기관을 위한 개인정보 비식별화에 적합한 기법들 세 가지를 설명하고자 한다.

1. 속성 값 삭제 기본[5]

속성 값 삭제는 원시 데이터에서 민감한 속성 값(주민등록번호, 나이, 성명, 주소 등)과 같이 개인 식별이 가능한 항목을 단순하게 제거하는 방법이다. 이 방법을 적용하였을 때 남아 있는 정보가 그 자체로 분석의 유효성을 가져야하고 또한 개인을 식별할 수 없어야한다. 그리고 인터넷 등에 공개되어 있는 정보 등과 결합하여 개인을 식별할 수 없어야 한다. NIST[4], IHE[3], HIPAA[5]의 경우 (그림 3)과 같이 속성 값 제거를 이용한 비식별을 위해 반드시 제거해야할 대상 정보에 대한 목록을 제공하여 직접 식별자들을 제거하도록 하고 있다. (그림 4)는 속성 값 삭제에 대한 예시를 보여주는 그림이다.



(2)(i) The following identifiers of the individual or of relatives, employers, or household members of the individual, are removed:

(A) Names	
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000	
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older	
(D) Telephone numbers	(L) Vehicle identifiers and serial numbers, including license plate numbers
(E) Fax numbers	(M) Device identifiers and serial numbers
(F) Email addresses	(N) Web Universal Resource Locators (URLs)
(G) Social security numbers	(O) Internet Protocol (IP) addresses
(H) Medical record numbers	(P) Biometric identifiers, including finger and voice prints
(I) Health plan beneficiary numbers	(Q) Full-face photographs and any comparable images
(J) Account numbers	(R) Any other unique identifying number, characteristic, or code, except as permitted by paragraph (c) of this section; and
(K) Certificate/license numbers	

(ii) The covered entity does not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information.

그림 3. HIPAA 세이프 하버 방법[5]

나이	직업	결혼	수입	병명
27	학생	미혼	400000	심장병
33	회사원	기혼	3000000	간암
42	회사원	미혼	2700000	위궤양
41	자영업	기혼	4300000	당뇨
24	학생	기혼	300000	간암

↓

직업	결혼	수입	병명
학생	미혼	400000	심장병
회사원	기혼	3000000	간암
회사원	미혼	2700000	위궤양
자영업	기혼	4300000	당뇨
학생	기혼	300000	간암

그림 4. 속성 값 삭제 예시

2. 속성 값 부분 삭제

속성 값 부분 삭제는 속성 값 삭제와 같이 민감한 속성에 대하여 전체를 삭제하는 것이 아닌 해당 속성의 일부 값만을 삭제하여 대표성을 가진 값으로 보이도록 하는 방법이다. 속성 값 부분 삭제의 경우, 다음에 설명하는 범주화와 유사한 경우가 발생할 수 있다. 그러나 범주화의 경우 주로 수치 데이터에 적용하는 경우가 일반적인데 반하여 속성 값 부분 삭제의 경우 수치 데이터를 포함하여 텍스트 데이터 등에도 폭넓게 활용이 가능하다. (그림 5)는 속성 값 부분 삭제에 대한 예시를 보여준다.

나이	성별	우편번호	질병
17	여	00000	당뇨
21	남	00001	간암
36	여	10000	화상
91	남	10001	골절

↓

나이	성별	우편번호	질병
	여	00000	당뇨
21	남	00001	간암
36	여		화상
	남		골절

그림 5. 속성 값 부분 삭제 예시

3. 범주화 기본[1]

범주화는 은폐화 방법이라고도 하며, 명확한 값을 숨기기 위하여 데이터를 평균 또는 범주의 값으로 변환하는 방식이다. 단, 데이터의 평균이나 범주로 전체를 표현할 경우 특정 속성을 지닌 개인으로 구성된 단체의 속성 정보 공개는 그 집단에 속한 개인의 정보를 공개하는 것과 같은 결과를 나타낸다. 따라서 이 경우에는 범주화를 비식별화 처리로 볼 수 없다. 예를 들어 특정 희귀병으로 구성된 집단에서 희귀병 속성에 대하여 범주화를 수행하는 경우 집단 내의 특정 개인이 희귀병 환자임을 공개하는 것과 같은 상황이 되고 이러한 경우 비식별가 되었다고 볼 수 없다. 위 (그림 6)은 주소에 대한 범주화에 대한 예시를 보여준다.

나이	주소	성별	병명
32	서울시 종로구 효자동	남	심장병
26	서울시 용산구 후암동	여	폐암
44	서울시 종로구 청운동	남	당뇨
30	서울시 영등포구 양평동	여	당뇨
53	서울시 구로구 구로동	여	심장병
48	서울시 용산구 후암동	남	간암
39	서울시 구로구 오류동	여	위염

↓

나이	주소	성별	병명
32	서울시 종로구	남	심장병
44	서울시 종로구	남	당뇨
30	서울시 영등포구	여	당뇨
53	서울시 구로구	여	심장병
39	서울시 구로구	여	위염
26	서울시 용산구	여	폐암
48	서울시 용산구	남	간암

그림 6. 범주화 예시

Ⅳ. 결론 및 향후 연구방향

본 논문은 지난 2015년 미래창조과학부에서 발표된 빅데이터 비식별화 기술 활용 안내서를 기반으로 영국의 ICO[2], 미국의 IHE[3], NIST[4], HIPPA[5]에서 발표한 의료기관의 개인정보보호 비식별화 기술에 관한 가이드들을 검토하여 국내 의료기관에서 활용할 수 있는 기술적인 방안과 상관도를 제시한 것이다. (표 1)에서 나타낸 바와 같이 특히 상관도가 6인 데이터값 삭제(속성 값 삭제와 속성 값 부분 삭제) 방법은 국내 의료기관에서도 가장 널리 활용될 수 있는 방안이며 그 외에도 범주화 방법은 상관도가 3으로 나타나 두 번째 대안으로 적용해 볼 수 있는 방법이다. 향후 연구방향으로는 앞서 제시한 정형화 데이터에 대한 대안들을 실제 실험을 통해 그 성능을 시험해 볼 예정이며, 아울러 비정형 데이터들에 대한 국내 적용방안들에 대해서도 제시해 보고자 한다.

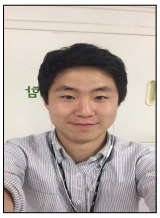
감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.B0713-15-0007, 의료정보 생애주기를 고려한 현장중심의 국제표준 스마트 의료 보안 플랫폼 개발).

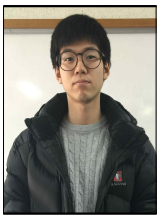
References

[1] 양현철, 김자영, 김진철, 김배현, 신신애, “빅데이터 비식별화 기술 활용 안내서 ver 1.0”, 미래창조과학부, 한국정보화진흥원, Vol. 1, 2015
[2] ICO (Information Commissioner’s Office, 영국), “Anonymisation: managing data protection risk code of practice”, 2012
[3] IHE IT Infrastructure Technical Committee, “IHE IT Infrastructure Handbook De-Identification”, Rev. 1.1, 2014
[4] Simson L. and Garfinkel, NIST IR 8053, “De-Identification of Personal Information”, U.S. Department of Commerce, 2015
[5] HIPAA Compliance Assistance, “Summary of the HIPAA Privacy Rule”, 2003
[6] Bernhard Riedl, Thomas Neubauer, and Gernot Goluch, “A secure architecture for the pseudonymization of medical data”, IEEE Computer Society, 2007
[7] 이창무, 오승교, 최덕재, “활성산초 측정 데이터를 위한 모바일기반의 U헬스 시스템 설계 및 구현, 스마트미디어저널, Vol. 1, No. 4, pp 59-71, 2012년 12월

저 자 소 개



여광수(학생회원)
2012년 한라대학교 컴퓨터공학과 입학
2016년 한라대학교 컴퓨터공학과 재학 중
<주관심분야 : 의료정보보안, 빅데이터>



김철중(학생회원)
2012년 한라대학교 컴퓨터공학과 입학
2016년 한라대학교 컴퓨터공학과 재학 중
<주관심분야 :의료정보보안>

이재현(정회원)
1989년 중앙대학교 컴퓨터공학과 학사 졸업.
2001년 중앙대학교 컴퓨터공학과 석사 졸업.
2007년 연세대학교 인지과학 박사 졸업.
2016년 현재 강릉원주대 정보기술공학과 재직 중
<주관심분야 : 의료정보윤리, 의료정보보안 표준>

김순석(정회원)
1999년 한국정보보호진흥원 기술기준팀 연구원.
2003년 중앙대학교 컴퓨터공학과 박사 졸업.
2016년 현재 한라대학교 컴퓨터공학과 재직 중
<주관심분야 : 의료정보보안, 표준>