http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.24, No.1, February 2019, pp.121-137
https://doi.org/10.7838/jsebs.2019.24.1.121

의료 비정형 텍스트 비식별화 및 속성기반 유용도 측정 기법

De-identifying Unstructured Medical Text and Attribute-based Utility Measurement

노 건(Gun Ro)*, 전종훈(Jonghoon Chun)**

초 록

비식별화는 데이터셋으로부터 개인정보를 제거함으로써 개인을 식별할 수 없도록 하는 방법으로, 정보를 수집, 가공, 저장, 배포하는 과정에서 발생할 수 있는 개인정보 노출 위험도를 낮추기 위해 사용한다. 그간 비식별화와 관련된 알고리즘, 모델 등의 관점에서 많은 연구가 이루어졌지만, 대부분은 정형 데이터를 대상으로 하는 제한적인 연구로, 비정형 데이터에 대한 고려는 상대적으로 많지 않은 실정이다. 특히 비정형 텍스트가 빈번히 사용되는 의료 분야의 경우에서는 개인 식별 정보들을 단순 제거함으로써 개인정보 노출 위험도는 낮추지만, 그에 따른 데이터 활용성이 떨어지는 점을 감수하는 실정이다. 본 연구는 개인정보 보호 이슈가 가장 중요하고 따라서 비식별화가 활발하게 연구되고 있는 의료분야 데이터 중 비정형 텍스트를 대상으로 k-익명성 보호모델을 적용한 비식별화 수행 방안을 제시하고, 비식별화 결과에 대한 새로운 유용도 측정 기법을 제안하여 이를 통해 직관적으로 데이터 활용성을 판단할 수 있도록 하는 것을 목표로 한다. 따라서 본 연구의 결과물이 의료 분야뿐만 아니라 비정형 텍스트가 활용되는 모든 산업 분야에서 활용될 경우, 개인 식별 정보가 포함된 비정형 텍스트의 활용도를 향상시킬 수 있을 것으로 기대한다.

ABSTRACT

De-identification is a method by which the remaining information can not be referred to a specific individual by removing the personal information from the data set. As a result, de-identification can lower the exposure risk of personal information that may occur in the process of collecting, processing, storing and distributing information. Although there have been many studies in de-identification algorithms, protection models, and etc., most of them are limited to structured data, and there are relatively few considerations on de-identification of unstructured data. Especially, in the medical field where the unstructured text is frequently used, many people simply remove all personally identifiable information in order to lower the exposure risk of personal information, while admitting the fact that

* First Author, Department of Computer Engineering, Myongji University, Yongin, Korea (laylow861@gmail.com)
** Corresponding Author, Department of Data Technology, School of Software Convergence, Myongji University, Seoul, Korea(jchun@mju.ac.kr)
Received: 2019-02-13, Review completed: 2019-02-19, Accepted: 2019-02-25

the data utility is lowered accordingly. This study proposes a new method to perform de-identification by applying the k-anonymity protection model targeting unstructured text in the medical field in which de-identification is mandatory because privacy protection issues are more critical in comparison to other fields. Also, the goal of this study is to propose a new utility metric so that people can comprehend de-identified data set utility intuitively. Therefore, if the result of this research is applied to various industrial fields where unstructured text is used, we expect that we can increase the utility of the unstructured text which contains personal information.

키워드 : 비식별화, k-익명성, 유용도 측정 기법, 의료 비정형 텍스트 비식별화
De-Identification, k-Anonymity, Utility Measurement, Unstructured Text De-Identification

1. 연구 개요

비식별화(De-identification)는 데이터셋으로부터 개인 식별 정보를 제거함으로써 남은 정보가 특정 개인을 식별할 수 없도록 하는 방법으로, 이를 통해 정보를 수집, 가공, 저장, 배포하는 과정에서 발생할 수 있는 개인정보 노출 위험성을 낮출 수 있다[8]. 비식별화라는 개념이 중요해진 것은 빅데이터라고 불릴 만큼 방대한 양의 데이터에 대한 분석이 필요해지면서 개인정보 보호에 대한 이슈가 점점 대두되어가고 있기 때문이다. 빅데이터 시대의 정보 프라이버시 위험과 정책에 관한 실증 연구에서는 이러한 프라이버시 보호 정책과 개인정보 제공 여부와의 상관성 연구를 통해 정보 프라이버시 위험을 낮추는 것이 매우 중요함을 연구한 바 있다[13]. 이와 같이 개인정보 보호 이슈는 빅데이터를 다루는 거의 모든 산업 분야에서 대두되고 있으며, 특히 많은 양의 환자 정보를 다루는 의료분야에서는 더욱 민감하게 다루어지고 있다. 미국 Health Insurance Portability and Accountability Act(HIPAA)의 개인정보 보호법은 의료 데이터에 포함되어 있는 개인을 식별할 수 있는 의료정보 18가지의 항목을 지정하고 이를 PHI(Protected Health Information)라고 칭하여 보호 대상으로 삼는다[12]. 또한 PHI를 보호함과 동시에 데이터의 유용성을 확보하기 위해 비식별화를 도입, 명시적으로 두 가지 방법(전문가에 의한 결정 또는 개인 식별자를 포함한 다른 정보와 결합되었을 때 개인을 식별할 수 있는 준식별자의 제거)에 의해서만 비식별화를 수행할 것을 규제하고 있다[12]. 관련하여 그간 비식별화 알고리즘, 비식별화 모델 등의 관점에서 많은 연구가 이루어졌지만, 대부분의 연구는 정형 데이터를 대상으로 제한적으로 진행되어 왔으며, 비정형 데이터에 대한 비식별화 연구는 상대적으로 많지 않은 실정이다.

본 연구는 비식별화가 가장 활발하게 연구되고 있는 의료분야에서 발생되는 데이터 중 의사 진단 노트, 간호일지 등 비정형 텍스트를 대상으로 k-익명성 보호모델을 적용한 비식별화를 수행하는 방안을 제시한다. 기존의 의료분야의 비정형 텍스트를 대상으로 수행한 비식별화 연구들은 HIPPA에서 규정한 방법 중 두 번째 방법인 '모든 PHI 항목들에 대한 단순 제거'를 적용한 것으로, 개인정보 노출 위험도는 최대한

낮출 수 있지만, 그만큼 수행된 결과에 대한 데이터 유용도는 현저히 낮을 가능성이 높다. 본 연구에서 제시하는 k-익명성 보호모델 적용을 통한 비식별화 방법은 HIPPA에서 규제하는 비식별화 방법 중 '전문가에 의한 결정'에 해당된다. 전문가의 결정에 의해 일부 PHI 항목들에 대해 단순 제거가 아닌 일반화를 거쳐 비식별화를 수행하게 되므로, 기존의 방식들과 비교하였을 때 데이터 활용 목적에 부합하는 비식별화를 수행할 수 있으므로 데이터 유용도를 높일 수 있는 방안이라고 판단한다. 또한, 본 연구에서 활용하는 k-익명성 보호모델과, 이에 따른 데이터 유용도 측정 방식들은 비정형 데이터가 아닌 정형 데이터를 대상으로 수행되는 방법이다. 따라서 이를 비정형 텍스트에 그대로 적용할 수 없다. 게다가, 기존 유용도 측정 방식들은 수치만 제공할 뿐, 이를 통해 비식별화된 결과가 어느 정도의 유용도를 가지고 있는지 사용자가 직관적으로 판단하기 어렵다는 단점이 있다. 따라서 본 연구는 비정형 텍스트를 대상으로 k-익명성을 적용한 비식별화 수행 결과에 대해 새로운 유용도 측정 기법을 제시하고, 측정 결과를 통해 직관적인 유용도 판단이 가능하도록 하는 것을 목적으로 한다.

2. 비식별화(De-identification)

2.1 비식별화

비식별화(De-identification)는 데이터셋으로부터 개인정보를 제거함으로써 남은 정보가 특정 개인을 식별할 수 없도록 하는 방법이다. 비식별화를 통해 정보를 수집, 가공, 저장, 배포하는 과정에서 발생할 수 있는 개인정보 노출 위험도를 낮출 수 있으며, 일반적으로는 정형 데이터셋을 대상으로 수행된다. 비식별화는 데이터셋에서 개인 식별 정보를 대상으로 수행되는데, 개인 식별 정보는 식별자, 준식별자, 민감 정보로 크게 나뉜다. 식별자(Direct Identifier)는 항목 하나만으로도 개인을 식별할 수 있는 정보를 의미하며, <Table 1>과 같은 병원의 환자 정보를 나타내는 데이터셋의 경우 환자 개인을 식별할 수 있는 환자명 항목이 식별자에 해당된다. 준식별자(Quasi-Identifier)는 식별자와 같이 항목 하나만으로 개인을 식별할 수는 없지만, 다른 데이터셋과의 결합 또는 배경지식 등을 통해 개인을 식별할 수 있는 정보를 의미하며, <Table 1>에서는 출생지, 출생 년도, 진단명이 준식별자에 해당된다. 일반적으로는 식별자를 제외한 모든 항목을 준식별자로 가정한다. 민감 정보(Sensitive Attribute)는 해당 정보의 소유자가 공개하기를 꺼려하는 정보, 즉 공개하기에는 민감한 정보를 의미하며, <Table 1>에서는 진단명이 이에 해당될 수 있다. 진단명은 준식별자로 분류될 수도 있고, 민감 정보로 분류될 수도 있는데, 이는 데이터 소유자의 판단 또는 데이터의 사용 목적에 따라 달라질 수 있다. 일반적으로 의료 분야에서 진단명은 공개되지 않았으면 하는 정보, 즉 민감 정보에 해당되어, 식별자와 같이 공개 대상에서 제외되는 것이 보통이다. 다만 데이터 소유자로부터 승인을 받고 연구나 분석 목적으로 공개된 데이터셋의 경우에는 이러한 민감 정보들을 준식별자로 판단하는 경우도 있다.

데이터셋을 비식별화 하는 방법은 대상이 식별자, 준식별자, 또는 민감 정보이냐에 따라 다

〈Table 1〉 Structured Data Consists of Patients' Information from Hospital

Patient's name	Location of Birth	Year of Birth	Diagnosis
John	Seattle, WA	1970	Obesity
Mark	Los Angeles, CA	1973	Obesity
Jane	Las Vegas, NV	1974	Obesity
Kim	Seoul, Korea	1953	Chest Pain
Kaito	Tokyo, Japan	1958	Headache
Alicia	Singapore	1953	Abdominal Pain
Gabriel	Paris, France	1982	Hypertension
Antonio	Rome, Italy	1986	Peliosis Hepatis
Walter	Berlin, Germany	1984	Hypertension

른 방법을 적용한다. 예컨대, 식별자, 민감 정보의 경우, 값을 전부 가리는 Masking(예: 홍길동 → xxx), Hashing, 또는 가명으로 대체하는 Pseudonymization(예: 홍길동 → 이준현) 등이 있다. 준식별자의 경우, Micro Aggregation, Generalization of Categories, Data Suppression 등을 통해 준식별자에 대한 비식별화를 수행한다[14].

2.2 일반화(Generalization)

일반화는 값 자체의 정확도를 변형함으로써 일반화된 값으로 대체하는 방법이다[5]. 준식별자 값의 유형에 따라 1~10과 같은 범위 또는 상위 분류로의 값 변경 등을 통해 일반화를 수행할 수 있다. 예를 들어, <Table 1>에서 John에 대한 정보의 경우, 출생지 'Seattle, WA'는 도시명이므로 상위 분류인 'US'로 일반화할 수 있으며, 출생년도인 '1970'은 '1970s' 또는 '1961~1970'과 같은 형태로 일반화할 수 있다. 이러한 일반화의 규칙을 준식별자 별로 정의한 것을 일반화 체계(VGH: Value Generalization Hierarchy)라고 한다[15]. <Table 1>의 준식별자인 출생지에 대해 일반화 체계는 <Table 2>와 같이 구성될 수 있다. <Table 1>이 나타내는 모든 출생지 값을 도시, 국가, 대륙, *로 총 4단계로 표현이 가능하며, 단계가 높아질수록 값의 정확도가 낮아지고, 더 많은 값들이 동일 또는 유사하게 변경되므로, 개인을 식별할 수 있는 확률도 그만큼 낮아지게 된다. 예를 들어, 0단계일 경우 'Las Vegas, NV'라는 출생지를 가진 환자는 Jane 한 명뿐이지만, 2단계일 경우에는 'North America' 출생지를 가진 환자가 3명으로 증가함으로써 Jane을 식별할 수 있는 확률이 1/3로 줄어들게 된다. 4단계의 일반화를 적용할 경우에는 모든 출생지 값이 '*'로 일반화되므로, Jane을 식별할 수 있는 확률이 1/9로 더욱 줄어들게 된다. 이처럼 적용하는 일반화 단계를 높일수록 개인을 식별할 수 있는 확률은 낮아지며, 반대로 일반화 단계를 낮출수록 개인 식별 확률은 높아진다. 따라서 일반화를 통한 비식별화 수행 시에는 일반화의 어느 단계를 적용하느냐는 비식별화 결과의 개인 식별 확률에 큰 영향을 미친다.

〈Table 2〉 Value Generalization Hierarchy of Location of Birth from 〈Table 1〉

Level 0	Level 1	Level 2	Level 3
Seattle, WA	US	North America	*
Los Angeles, CA	US	North America	*
Las Vegas, NV	US	North America	*
Seoul, Korea	Korea	Asia	*
Tokyo, Japan	Japan	Asia	*
Singapore	Singapore	Asia	*
Paris, France	France	Europe	*
Rome, Italy	Italy	Europe	*
Berlin, Germany	Germany	Europe	*

2.3 k-익명성 보호 모델(Protection Model)

k-익명성 보호 모델은 Latanya Sweeney가 제안한 개념으로, 공개된 데이터셋에서 각각의 레코드와 동일한 준식별자 값을 가지고 있는 레코드의 개수, 즉 동질 집합의 크기가 적어도 k 이상이 되도록 일반화를 통해 강제함으로써, 특정 개인을 식별할 수 없도록 하는 방법이다[16]. 예를 들어, 어떤 의사가 <Table 1>의 데이터셋에 대해 k-익명성(k = 3)이 수행된 결과를 통해 출생지, 출생년도, 진단명을 확인하고자 한다면, 환자명은 식별자로 지정하고, 확인하고자 하는 항목인 출생지, 출생년도, 진단명 항목은 준식별자로 지정한다. k 값이 적용되는 모든 준식별자는 <Table 2>와 같이 VGH가 준식별자별로 정의되어야 하며, 이를 통해 k-익명성(k = 3)을 수행한 결과는 <Table 3>과 같이 나타날 수 있다.

의사가 <Table 3>을 통해 어느 한 명의 환자를 식별하고 싶어도 그와 동일한 레코드가 3개 존재하기 때문에, <Table 1>과 비교하였을 때 개인을 식별할 수 있는 확률이 1/3로 낮아지게 된다. 즉, 동질 집합의 크기를 나타나는 k 값이

〈Table 3〉 Example of Applying k-anonymity(k = 3) to 〈Table 1〉

Patient's name	Location of Birth	Year of Birth	Diagnosis
*	North America	1970's	Nutrition and Metabolic Disease
*	North America	1970's	Nutrition and Metabolic Disease
*	North America	1970's	Nutrition and Metabolic Disease
*	Asia	1950's	Pain
*	Asia	1950's	Pain
*	Asia	1950's	Pain
*	Europe	1980's	Vascular Disease
*	Europe	1980's	Vascular Disease
*	Europe	1980's	Vascular Disease

커지면 커질수록 그만큼 개인을 식별할 수 있는 확률은 더욱 낮아지게 된다. 또한, 준식별자로 지정된 항목인 출생지, 출생년도, 진단명은 VGH 적용을 통해 k 값을 만족하는 선상에서 최소한의 일반화된 값으로 대체되기 때문에 단순 제거를 통한 비식별화 방법과 비교했을 때 상대적으로 데이터 유용도도 보존될 수 있다.

3. 필요성 및 유사 연구

본 연구는 비식별화 등의 소프트웨어 개발 및 평가 목적으로 사용되는 간호 일지(nursing notes) 비정형 텍스트를 대상으로 k-익명성 보호모델을 적용한 비식별화를 수행하는 방안을 제시하고, 수행 결과에 대한 새로운 유용도 측정 기법을 마련하여, 직관적인 유용도 판단이 가능하도록 한다.

비식별화는 많은 산업 분야에 적용되어 활용되고 있지만, 특히 의료 분야에서는 환자들로부터 발생하는 데이터를 다루기 때문에 개인정보 보호를 위해 필수 불가결한 요소로 자리잡았다. 일반적으로, 환자 정보 보호를 위하여 중요한 정보들은 모두 식별자, 민감 정보로 분류하고, k-익명성 적용 시에 타 분야에 비해 상대적으로 높은 k 값을 적용하여, 개인 식별 확률을 최대한 낮추는 것이 보통이다. 하지만, 정밀 의료를 통해 환자들의 증상, 질병들을 더 정확하게 분석하고 이에 대한 해결책 마련을 위해 더 많은 양, 더 자세한 정보를 가진 데이터가 필요한 의사, 연구진 입장에서는 이러한 비식별화 방법을 달가워하지 않는 것이 현실이다. 따라서, 비식별화를 통해 개인 식별 확률은 일정 수준 이하로 낮게 유지하는 동시에, 데이터의 활용, 연구를 위해 충분한 데이터 유용도도 보장할 수 있도록 하는 비식별화 방법의 필요성이 대두되었다.

HIPAA에서는 명시적으로 첫 번째 방법인 전문가에 의한 결정, 두 번째 방법인 개인 식별자를 포함한 다른 정보와 결합되었을 때 개인을 식별할 수 있는 준식별자의 제거에 의해서만 비식별화를 수행할 것을 규제하고 있다. 첫 번째 방식은 통계적, 과학적인 원리 즉, k-익명성과 같은 보호 모델 알고리즘을 적용한 방법이며, 두 번째 방식은 'Safe Harbor'로 일컫는 방법으로, HIPPA에서 명시한 18가지 PHI 항목에 해당에 해당하는 모든 값을 제거하는 방법이다[12].

HIPAA에서 명시한 비식별화 방법에 의거하여, 의료분야에서 발생하는 비정형 텍스트의 비식별화와 관련된 연구로는 free-text로 이루어진 의료 기록 데이터를 대상으로 "gold standard corpus"를 구축하고, 이를 통해 자동화된 비식별화를 수행하는 연구[11]가 있으며, 타 데이터와의 결합을 통해 발생할 수 있는 재식별을 고려, 장기간에 걸쳐 생성되는 임상실험 텍스트를 대상으로 한 비식별화 자동화 시스템 연구[6]가 있다. 두 연구 모두 HIPAA에서 지정한 PHI 항목들 전체 또는 일부를 비식별화 대상으로 하고, 대상 데이터로부터 지정된 PHI 항목들의 추출 정확도를 정확도, 재현율, 그리고 F1-점수를 통해 측정하는 방식으로 자신들의 연구 결과의 우수성을 입증하였다. 또한, 두 연구 모두 HIPPA에서 명시한 "Safe Harbor" 방법, 18가지 PHI 항목들에 대한 제거 또는 가명화를 통한 비식별화를 수행하기 때문에 지정한 항목들을 정확하게 추출하고 나면 연구의 목표가 달성된 것이라고 봐도 무방하다. 하지만, 이러한 방법은 PHI 항목들에 대해 전부 제거되거나 가명화 되었기 때문에 개

인을 식별할 수 있는 확률은 낮아질 수 있지만, 연구나 분석을 위해 사용하기에는 유용도가 떨어질 가능성이 높다. 따라서 비식별화된 데이터의 유용도를 유지하면서 동시에 개인정보 식별 위험도는 낮출 수 있는 비식별화 방법이 필요하다 하겠다.

본 연구는 PHI 항목들을 대상으로 한 비식별화를 수행하되, HIPAA에서 규정한 첫 번째 방법인 전문가에 의한 결정 방법을 적용, PHI 항목들에 대한 무조건적인 제거가 아닌 k-익명성 비식별화 보호 모델을 사용함으로써, 데이터 유용도의 훼손을 최소화하는 비식별화 방안을 제시하고, 이를 정량적으로 측정하여 달성도를 판단하기 위한 체계적인 방안을 제시하는데 그 목적이 있다.

4. 연구 실험 및 결과

4.1 실험 환경

본 연구 실험은 CPU i7 2.8GHz, RAM 16GB, HDD SSD 512GB 사양을 갖춘 맥북 프로를 기반으로 수행하며, 실험 장비에 mySQL을 설치하여 이에 비식별화 결과를 저장한다. 또한, 모든 실험 코드는 JAVA8을 기준으로 작성하며, 비식별화 수행을 위해 ARX 라이브러리를 활용한다.

4.2 실험 대상 데이터셋

NER(Named Entity Recognizer)을 활용한 식별자, 준식별자의 추출에 대한 precision과 recall 측정 실험 연구에서는 HIPPA에서 지정한 18가지 PHI 항목들을 식별자, 준식별자로 분류하고, 2개의 데이터셋인 Reuters-21578[10]과 Gold Standard Corpus로부터 이 항목들을 Stanford Classifier[4]를 활용하여 얼마나 정확하게 추출하는지를 측정하였다.

<Table 4>는 위 연구에서 사용한 데이터셋에 대한 간략한 특징을 보여준다. Reuters-21578은 IR 분야 및 NER에서 많이 사용되는 공개 데이터셋이며. 21,578건의 뉴스 기사로 이루어져 있다, 각 뉴스 기사는 제목, 본문, 날짜 항목들을 가지고 있지만, 본 연구는 이 중에 본문만을 대상으로 수행한다. 두 번째 데이터셋은 비식별화 자동화 연구[11]에 사용된 Gold Standard Corpus이다. 2,434건의 의료 간호일지로 구성되어 있으며. 각 간호일지에 포함된 환자 이름은 실제 이름과는 전혀 무관한 다른 값으로 대체되어, 개인을 식별할 수 없는 상태로 비식별화된 데이터셋이다.

〈Table 4〉 Reuters-21578 and Gold Standard Corpus

	Reuters-21578	Gold Standard Corpus
Number of Unstructured Text	21578	2434
Description	News Articles	Nursing Notes
Field	Not in Particular	Medical
Data Structure	XML. each article has title, body, and date element	Not in Particular
Characteristic	Each article may have more than two individual names	Each nursing note is about one patient

본 연구는, 위 연구에서 사용한 식별자, 준식별자 추출 방법을 활용하여, Reuters-21578 데이터셋으로부터 추출된 준식별자들을 대상으로 비식별화를 수행하고, 수행결과에 대한 유용도를 측정한다.

4.3 비식별화

비식별화 수행에는 오픈소스로 공개되어 있는 ARX: Data Anonymization Tool에서 제공하는 라이브러리를 활용한다[3]. ARX를 통해서 비식별화 수행 시에 k-익명성을 포함한 다양한 보호 모델은 물론 사용자가 원하는 형태의 일반화 체계를 적용할 수 있어, 본 연구의 비식별화 실험에도 ARX를 활용한다.

4.3.1 일반화 체계(VGH: Value Generalization Hierarchy) 구성

각 비정형 텍스트로부터 식별자 및 준식별자로 추출되어 저장된 단어들을 대상으로 비식별화를 수행하되, 준식별자는 각 값들이 가지고 있는 특성을 분석하여, 이에 맞는 일반화 체계를 적용하여야 한다. <Table 5>는 Reuters-21578 데이터셋으로부터 Location으로 분류된 단어들에 대한 일반화 체계를 직접 구성한 결과의 일부이다. 사람의 나이처럼 숫자로 이루어진 특정 범위로 분류가 가능한 준식별자들은 물론, <Table 5>와 같이 카테고리화 해야 하는 준식별자들도 ARX 기능을 통해 반자동 VGH 생성이 가능하지만, 카테고리형 준식별자로부터 생성되는 VGH는 사용자가 원하는 형태로 구성되어 있지 않은 경우가 대부분이어서, <Table 5>와 같은 형태로의 자동 생성은 불가능하다.

〈Table 5〉 VGH of Location

Level 0	Level 1	Level 2	Level 3
Manhattan	US	America	*
Gallatin	US	America	*
Baghdad	Iraq	Middle East	*
Atlanta	US	America	*
Ottawa	US	America	*
California	US	America	*
Texas	US	America	*
Connecticut	US	America	*
sub-Saharan	Middle East	Middle East	*
Washington	US	America	*
Caracas	Venezuela	America	*
Utah	US	America	*
Pennsylvania	US	America	*
Moscow	Russia	Asia	*
Alabama	US	America	*
Greenwich	UK	Europe	*
Chicago	US	America	*
Greenwich	UK	Europe	*

<Table 5>는 총 4단계로 일반화 단계를 구성하였으며, 0단계는 원본 값인 도시 또는 지역명, 1단계에서는 이를 국가명, 2단계에서는 대륙명, 3단계에서는 *로 일반화되도록 하였다. 단, 모든 단어들에 대해 이 일반화 단계를 모두 똑같이 적용할 수는 없다. 예를 들어, <Figure 8>의 추출된 단어가 sub-Saharan과 같이 대륙의 일부를 일컫는 지역명은 0단계에서 이미 대륙명이므로, 다른 단어들과 비교하였을 때 부각되어 보일 수밖에 없다. 이런 경우에는 단계가 올라갈수록 최대한 다른 단어들과 동일한 값을 가질 수 있도록 구성한다. sub-saharan의 경우에는 1단계부터 Middle East라는 값을 지정, 2단계에서부터 다른 2개의 단어와 동일한 값을

```
public static AttributeType.Hierarchy.DefaultHierarchy getLocationHierarchy() {
    AttributeType.Hierarchy.DefaultHierarchy locationHierarchy = AttributeType.Hierarchy.create();

    locationHierarchy.add("Manhattan", "US",  "America", "*");
    locationHierarchy.add("Gallatin",  "US",  "America", "*");
    locationHierarchy.add("Malta",     "US",  "America", "*");
    locationHierarchy.add("Atlanta",   "US",  "America", "*");
    locationHierarchy.add("Ottawa",    "US",  "America", "*");
    locationHierarchy.add("California","US",  "America", "*");
    locationHierarchy.add("Texas", "US",  "America", "*");
    locationHierarchy.add("Connecticut", "US",  "America", "*");
    locationHierarchy.add("sub-Saharan", "Middle East", "Middle East", "*");
    locationHierarchy.add("Washington",  "US",  "America", "*");
    locationHierarchy.add("Maine", "US",  "America", "*");
    locationHierarchy.add("Utah",  "US",  "America", "*");
    locationHierarchy.add("Pennsylvania",  "US",  "America", "*");
    locationHierarchy.add("Indiana", "US",  "America", "*");
    locationHierarchy.add("Alabama", "US",  "America", "*");
    locationHierarchy.add("Ohio",  "US",  "America", "*");
    locationHierarchy.add("Chicago", "US",  "America", "*");
    locationHierarchy.add("Greenwich", "UK",  "Europe",  "*");
    locationHierarchy.add("Michigan",  "US",  "America", "*");
    locationHierarchy.add("Missouri",  "US",  "America", "*");
```

〈Figure 1〉 Applying VGH of Location on ARX

가질 수 있도록 구성하였다. ARX 라이브러리에서는 구성된 VGH를 2차원 배열 형태로 적용하도록 되어 있으며, 따라서 <Figure 1>과 같은 형태로 코드가 수행되게 된다.

4.3.2 비식별화 수행 결과

Reuters-21578 데이터셋 중 무작위로 50건의 뉴스 기사로부터 사람 이름과 장소 단어를 추출하고, 이 추출 결과를 대상으로 <Figure 1>의 일반화 체계를 적용하여 k-익명성(k = 3)을 수행하였다. 50건의 뉴스는 대부분 금융, 경제에 대한 내용으로 구성되어 있고, 이로부터 준식별자로 추출된 355개의 단어에 대해 총 374ms의 비식별화 수행 시간이 소요되었으며, 약 30.78 MB의 메모리를 사용하였다. <Table 6>은 이 수행 결과의 일부를 나타낸다.

사람 이름은 식별자로 지정하였기 때문에 전부 '*'로 대체되었으며, Location 값에 대해서만 k-익명성을 통한 일반화가 적용되었다. 앞서 정의한 일반화 체계에서 1단계가 적용된 값이며, 일

〈Table 6〉 De-identification Result of PERSON and LOCATION

PERSON	LOCATION
*	U.S.
*	*
*	*
*	U.S.
*	U.S.
*	U.S.
*	Africa
*	U.S.
*	U.S.
*	U.S.
*	*
*	Africa
*	Africa
*	U.S.
*	U.S.
*	U.S.
*	U.S.
*	*

부가 대륙명인 Africa로 되어 있는데, 이는 일반화 체계 구성 시에 Africa에 속하는 단어들이 0단계에서 이미 국가명으로 지정되어 있기 때문이다.

4.4 유용도 측정

데이터 유용도(이하 유용도)는 비식별화가 적용된 데이터셋이 내포하고 있는 정보가 얼마나 유용한지를 판단하는 기준을 말한다. 일반적으로는, k-익명성의 경우 k 값을 높이면 높일수록 개인을 식별할 수 있는 확률은 낮아지지만, 반대로 그만큼 유용도는 낮아질 수 있다. 따라서, 비식별화 수행 결과를 통해 개인을 식별할 수 있는 확률이 높지 않으면서 동시에 유용도 손실을 최소화 하는 것이 본 연구의 목적이며, 이를 위해 비정형 텍스트 대상으로 한 비식별화 결과에 적합한 유용도 측정 방식을 제안한다.

4.4.1 기존 유용도 측정 방식 적용에 대한 한계

본 연구의 목적은 정형 데이터에 맞춰진 k-익명성 보호 모델을 적용하여 비식별화를 수행하고, 이 결과에 대한 유용도 수치가 보호 모델을 적용하지 않고 단순 제거하는 방법보다 높게 도출되도록 하는 것이다. 하지만 k-익명성은 정형 데이터에 적용되는 보호모델이므로, k-익명성을 적용하여 수행한 비식별화 결과의 유용도 측정 방식 또한 정형 데이터에 사용되는 방식을 고려할 필요가 있다. 대표적으로 알려진 유용도 측정 방식으로는 일반화에 의한 정보 손실 측정(Generalization Information Loss), 분별력 측정(Discernibility Metric, DM)[15], 평균 동질집합 크기(Average Equivalence Class Size, CAVG) 3가지가 있다. 이 방식들은 모두 정형화된 데이터를 대상으로 하며, <Table 1>의 병원이 보유한 환자 정보와 같이 하나의 튜플이 한 명의 개인을 나타내는 것을 전제로 한다. 하지만, 비정형 텍스트로부터 추출된 식별자 및 준식별자들을 대상으로 비식별화가 수행된 결과는 위와 같이 하나의 튜플, 즉 하나의 뉴스 기사 또는 간호 일지가 한 명의 개인을 나타낸다고 보장할 수 없다. 따라서, 본 연구는 기존의 유용도 측정 방식을 보완, 비정형 텍스트를 대상으로 한 비식별화에 부합하는 새로운 측정 방식을 제안한다.

4.4.2 분별력(Discernibility Metric) 측정 방식

앞서 소개한 유용도 측정 방식 3가지 중 일반화에 의한 정보 손실 측정 방식은 비식별화 수행시 적용한 일반화 체계(VGH)를 기준으로 수행된 결과 값과의 단계 차이를 계산하기 때문에, 사용한 보호 모델은 물론 일반화 체계에 따라서도 유용도 수치가 매우 달라질 수 있다. 평균 동질집합 크기 방식은 동질 집합 크기가 클수록 유용도가 낮은 것으로 판단하는 방법으로 일반적으로 많이 사용되지만, 그만큼 보호 모델의 유무에 관계없이 적용되므로, 데이터 특성에 맞는 유용도 수치를 측정하기 어렵다. 반면에 분별력 측정 방식은 반드시 k 값을 적용한 비식별화 결과에 대해서만 측정이 가능하고, 유용도 수치에 대한 변수도 k만 존재하기 때문에, 본 연구는 분별력 측정 방식을 선택하여 보완한다.

기존의 분별력 측정(Discernibility Metric, DM)의 측정방식은 아래와 같다.

$$DM(T^*) = \sum_{\forall EQ s.t. |EQ| \geq k} |EQ|^2 + \sum_{\forall EQ s.t. |EQ| < k} |T| \cdot |EQ| \quad (1)$$

원본데이터 T에 대해 비식별화를 수행한 데이터 T^*의 분별력 $DM(T^*)$이 최종 결과 값이 되며, 기본적으로 동질 집합 크기가 k보다 크거나 같은 경우에는 각 동질집합(EQ) 크기의 제곱을 모두 더한 값이다. 하지만, 동질집합 크기가 k보다 작은 경우 각 동질 집합 크기에 원본 데이터 크기를 곱한 값을 더한다. 즉, 레코드가 포함된 동질 집합의 크기가 클수록 제곱 값이 더 많이 더해지게 되므로, 전체 DM 값이 커지게 되며, 결과적으로 동일한 값들을 가진 레코드들에 대해 페널티를 높게 부과하는 방식이다. 따라서, 결과 값이 낮을수록 유용도가 높다고 판단할 수 있다.

〈Table 7〉 Example of De-identified Result Apply k-Anonymity (k=3)

ID	PERSON	LOCATION	DATE
1	***	North America	1900's
2	***	North America	1900's
3	***	North America	1900's
4	***	Asia	1950's
5	***	Asia	1950's
6	***	Asia	1950's
7	***	Europe	1980's
8	***	Europe	1980's

<Table 7>에 나타난 비식별화 결과를 대상으로 위의 분별력 측정을 수행한다면, k > = 3 조건을 만족하는 두 동질 집합에 대해서는 제곱값이 더해지며, 이 조건을 만족하지 못하는 동질집합(Location = Europe, Date = 1980's)은 원본 데이터의 크기에 동질 집합 크기를 곱한 값이 더해진다. 따라서, $DM(T^*) = 3^2+3^2+8\times2 = 34$와 같은 결과를 갖게 된다.

일반적으로 비식별화 수행 시, k = 3만을 적용하여 k-익명성을 수행하게 되면, <Table 7>과 같은 결과가 도출되지 않는다. 그 이유는 k = 3이라는 것은 모든 동질 집합의 크기가 최소 3이라는 의미인데, <Table 7>의 id가 7, 8인 튜플은 동질집합의 크기가 2로, k = 3을 만족하지 않는다. 따라서, 모든 튜플이 k = 3을 만족할 수 있으려면, 앞서 정의한 일반화 체계를 적용할 경우 모든 Location 값이 '*'로 대체될 것이다. 모든 Location 값이 '*'로 대체되면 동질집합 크기가 8로 커지면서 k = 3을 만족하게 되지만, 반대로 모두 '*'로 대체되었기 때문에, 이에 대한 유용도는 떨어진다. 이처럼 모든 튜플에 대해 k 값을 충족시키기 위해서 나머지 6개의 튜플에 대한 유용도도 포기하는 것은 바람직하지 않을 수 있다. 그러므로, 이런 경우에는 id = 7, 8에 해당하는 튜플을 비공개 처리하고, 나머지 6개의 튜플에 대해서만 비식별화를 수행하는 것이 데이터 유용도를 향상시킬 수 있는 방법이다. 일반적으로는 전체 데이터 수의 %에 해당하는 수만큼의 튜플에 대해서는 전부 비공개 처리, 즉 '*' 처리되고, 나머지 튜플에 대해 k-익명성이 수행된다. 이 때 비공개 처리되는 튜플의 최대 수용 가능 수치를 suppression limit이라고 하며, 이를 통해 비공개 처리된 레코드들을 suppressed record라고 칭한다. 예를 들어, 1,000개의 record를 가진 데이터셋이 존재하고 이에 대해 suppression limit이 1%라고 가정하면, k-익명성을 적용한 비식별화를 수행할 경우,

최대 10 개의 record 에 대해서는 k 값을 만족하지 못하더라도 이에 대한 일반화를 수행하지 않고 비공개 처리하게 된다.

4.4.3 속성 기반 유용도 측정 기법 (Attribute-based Utility Measurement)

정형데이터 비식별화 결과에 사용되는 일반적인 분별력 측정 방식은 Record-Oriented 즉, 레코드별 동질집합들을 기준으로 계산하는 방식이다. 하지만, 본 연구에 사용된 비정형 텍스트는 하나의 레코드가 반드시 한 명의 개인정보에 해당한다는 보장이 없으므로, 레코드별이 아닌 준식별자별로 동질 집합에 대한 DM 값 (Attribute-Oriented DM)을 계산하기로 한다. 이를 수식으로 표현하면 아래와 같다.

$$DM(T^*|Ai) = \sum_{\forall EQs.t.|EQ| \geq k} |EQ|^2 + \sum_{\forall EQs.t.|EQ| < k} |T_{|A_i|}| \cdot |EQ| \quad (2)$$

식 (2)처럼 각 준식별자에 대한 DM 값이 계산되면, 이에 대한 합을 계산하고, 이를 준식별자 수로 나누어 평균을 구한다.

$$\frac{1}{n}\sum_{1}^{n} DM(T^*|Ai) \quad (3)$$

또한 위의 일반적인 분별력 측정 방식을 통해 도출되는 측정 수치인 '34'만을 보고서는 어느 정도 유용도가 있는지 직관적으로 판단하기가 어렵다. 따라서, 결과 값에 대해 0과 1사이의 값으로의 매핑을 통해 변환함으로써, 결과 값을 봤을 때 어느 정도 유용도를 가지는지 직관적으로 판단할 수 있도록 한다.

최대값은 각 준식별자별로 발생할 수 있는 최대 DM 값을 의미한다. DM 값이 최대라는 의미는 각 준식별자 별로 모든 튜플 하나의 동질 집합에 속한다는 의미이므로,

$$DM(MAX|A_i) = (N_i)^2 \quad (4)$$
$$(*N_i = number of tuples \text{ for } each\ quasi\text{-}identifier)$$

와 같이 준식별자별 전체 레코드 수의 제곱이 최대 DM 값이 된다. 이때에는 준식별자별로 EQ 개수가 하나이면서, 크기가 N일 경우에 해당된다.

최소값은 준식별자별로 모든 튜플이 각각 구별될 수 있을 경우이므로, 식 (2)에서 모든 동질 집합이 k보다 작은 경우에 속하게 된다. 따라서,

$$DM(MIN|A_i) = N_i \times 1 = N_i \quad (5)$$

와 같은 결과를 얻게 된다. 이도 또한 최대값과 마찬가지로, 모든 준식별자에 대한 최소값이 더해져야한다.

위 두 최대값, 최소값 결과를 이용하여, 식 (2)의 값을 0과 1 사이의 값으로 정규화하면,

$$\sum_{i=1}^{n} DM(T^*|Ai)_{normalized} = \frac{\sum_{i=1}^{n} DM(T^*|A_i) - \sum_{i=1}^{n} N_i}{\sum_{i=1}^{n} N_i^2 - \sum_{i=1}^{n} N_i} = \sum_{i=1}^{n}\left(\frac{DM(T^*|A_i) - N_i}{N_i^2 - N_i}\right) \quad (6)$$

와 같은 결과를 얻게 되며, 위와 같이 계산된 값은 최소 0, 최대 1의 값을 갖게 된다.

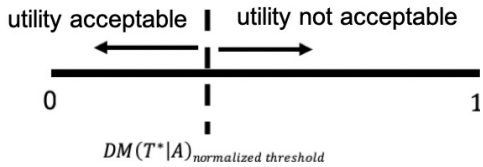
위의 정규화 과정을 통해 유용도 수치를 0과 1 사이의 값으로 변환하더라도, 이를 통해 직관

적으로 어느 정도의 유용도를 갖는지, 또는 과연 이 수치가 허용 가능한 범위 내의 수치인지 판단하기는 쉽지 않다. 따라서, 유용도 허용 한계값(threshold)을 계산하고, 이 한계값을 기준으로 유용도 수치에 대한 직관적인 판단이 가능하도록 하는 방안을 제시한다.

일반적으로 보호 모델을 적용하여 수행한 비식별화 결과에 대한 유용도는 비식별화 수행 시에 적용된 보호모델의 파라메터 값(예: k 값)은 물론 suppression limit에 매우 따라 달라질 수 있다. Khaled El Emam의 연구에서는 suppression limit을 1%, 5%, 10%로 지정하고 (k = 3, 5, … 22) 비식별화 수행시, suppression limit을 5%로 지정했을 때가 k 값에 관계없이 전체적으로 유용도가 가장 높은 것으로 보고되었다[2]. 따라서, 전체 데이터의 5%가 suppressed record 라고 가정한다면, 식 (7)을 통해 유용도 허용 한계값 $DM(T^*|Ai)_{threshold}$을 계산할 수 있다.

$$DM(T^*|Ai)_{threshold} = \sum_{\forall EQ s.t. |EQ| \geq k} |EQ|^2 + \sum_{\forall EQ s.t. |EQ| < k} |T_{|A_i|}| \cdot |EQ| = \sum_{\forall EQ s.t. |EQ| \geq k} |EQ|^2 + N_i \cdot 0.05N \quad (7)$$

suppression limit이 늘어나면 DM 값이 증가하고, 반대로 suppression limit이 줄어들면 DM 값이 감소하므로, 계산된 허용 한계 값인 DM(T*|Ai)threshold보다 작은 DM 값은 허용할 수 있는 값이라고 판단할 수 있다. 따라서, 이를 종합하여 정규화된 유용도 측정 값의 범위 및 이에 대한 최종 속성 기반 유용도 측정 계산식은 각각 <Figure 2> 및 식 (8)로 표현된다.



〈Figure 2〉 Domain of Normalized Utility Measurement with Normalized Threshold

$$\sum_{i=1}^{n} DM(T^*|Ai)_{normalized} = \sum_{i=1}^{n} \left(\frac{DM(T^*|A_i) - N_i}{N_i^2 - N_i}\right) \quad (8)$$
$$(*threshhold\ is\ when\ DM(T^*A_i)$$
$$= \sum_{\forall EQ s.t. |EQ| \geq k} |EQ|^2 + 0.05N_i)$$

4.1.4 유용도 측정 결과

<Table 7> 결과로부터 본 연구에서 제시한 속성 기반 유용도 측정 방식을 통해 유용도를 계산하면,

$$\sum_{1}^{n} DM(T^*|Ai) = 34 + 34 = 68$$

$$DM(MAX|A_i) = 8^2 = 64, DM(MIN|A_i) = 8$$

$$\sum_{i=1}^{n} DM(T^*|Ai)_{normalized}$$

$$= \left(\frac{34-8}{64-8} + \frac{34-8}{64-8}\right) \times \frac{1}{2}$$

$$= (0.46 + 0.46) \times \frac{1}{2} = 0.46$$

와 같이 각 준식별자 DM 값의 합은 68, 최대/최소 DM 값은 각각 64, 8로 계산되고, 이를 통해 속성 기반 유용도 측정값은 0.46으로 도출된다.

〈Table 8〉 Result of Applying Attribute-based Utility Measurement

number of news	number of words	DM(MAX|Ai)	DM(MIN|Ai)	Attribute-based Utility
50	355	53,824	232	0.1619

각 준식별자별 최소 허용 한계 값은 9+9+0.05×8 = 19이므로, 이에 대한 정규화된 최소 허용 한계 값은 0.357로 도출된다. 따라서 0.46의 수치는 최소 허용 한계 값보다 높게 측정되므로, 데이터 유용도가 낮은 것으로 판단할 수 있다. 기존의 DM 계산 방식을 통해 도출되었던 34라는 값과 비교하였을 때, 직관적으로 유용도를 판단할 수 있는 수치를 제공할 수 있다.

앞서 수행하였던 Reuters-21578 데이터셋으로부터 무작위로 50건의 뉴스 기사를 대상으로 사람 이름, 장소 단어를 추출하고 이에 대한 k-익명성(k = 3)을 수행한 결과인 <Table 5>를 대상으로 속성 기반 유용도 측정 방식을 통해 유용도를 계산하면, <Table 8>과 같은 결과를 얻는다.

5. 결론 및 향후 연구

5.1 결론

본 연구는 의료 비정형 텍스트를 대상으로 k-익명성 보호 모델을 적용한 비식별화 방안을 제시하였다. 기존 대부분의 의료 분야 비정형 텍스트에 대한 비식별화 연구에서는 모든 PHI에 대해 무조건적인 삭제 또는 치환을 수행하였다면, 본 연구는 정형 데이터에 사용되는 k-익명성 보호모델을 비정형 텍스트에 적용하여 준식별자들에 대한 비식별화를 수행, 기존의 의료 비정형 텍스트의 비식별화된 결과와 비교하였을 때 유용도를 높일 수 있는 방안을 제시하였다. 또한, 비식별화 결과에 적합한 새로운 유용도 측정 기법은 물론, 측정된 수치에 대해 직관적으로 유용도를 판단할 수 있는 방안을 제공함으로써, 비식별화에 대해 전문지식이 없는 일반 데이터 사용자들도 본 연구가 제시한 유용도 측정 방식 및 수치를 통해 비식별화된 데이터셋에 대해 손쉽게 유용도 판단이 가능하도록 하였다.

5.2 향후 연구

기존 정형 데이터셋에 대한 비식별화는 각 레코드가 하나의 개인을 의미하지만, 비정형 텍스트의 경우 비식별화를 대상으로 하는 문서 1건이 반드시 하나의 개인에 해당된다는 보장이 없다. 잘 알려진 대로 비정형 텍스트 문서에서 각 개인에 속하는 식별자, 준식별자들을 추출하고 이를 개인별로 그룹화 하는 것에는 한계가 존재한다. 향후에는 추출 대상인 비정형 텍스트에서 개체(entity) 추출 시, 주위 문맥을 파악하여 추출되는 개체를 개인별로 정확히 그룹화 할 수 있다면, 정형데이터와 더 근접하게 구조화를 시킬 수 있을 것으로 예상한다.

또한, 본 연구의 비식별화에 적용된 k 값은 일반적으로 많은 산업 분야에서 실제로 정형 데이터셋의 비식별화 수행 시에 적용하는 값으로, 개인정보 노출 위험도의 허용 범위를 만족하면

서 데이터 유용도도 일정 이상 제공할 수 있는 정도이다. 하지만, 정형 데이터셋과 본 연구에 사용된 비정형 텍스트로부터 구조화된 데이터셋은 본질적으로 다르기 때문에, 정형 데이터에 사용되는 k 값을 그대로 적용하더라도, 기대하는 만큼의 개인정보 노출 위험도를 낮추는 효과가 없을 수 있다. 예를 들어, k = 3이라고 할 경우, 일반적인 정형 데이터를 대상으로 적용한다면, 하나의 개인이 식별될 확률은 1/3이지만, 본 연구를 통해 구조화된 비정형 텍스트는 장소, 날짜 등에 속하는 값이 고유하다는 보장이 없기 때문에, k = 3이라는 값이 곧 개인이 식별될 확률을 1/k, 즉 1/3으로 만든다는 보장이 없다. 따라서, 확률적으로 정형 데이터에서 1/k의 개인 식별 확률을 만족하거나 유사한 근사 확률을 보장하려면, 비정형 데이터의 경우 문서 한건 당 평균 몇 명 정도의 개인이 포함되어 있는 환경에서 어느 정도의 k 값을 사용해서 비식별화를 수행해야 하는지 판단할 수 있는 기준을 마련하고 이를 실험적으로 입증해야 하는 과제가 남아있다.

마지막으로, 비정형 텍스트에 대한 비식별화를 수행한다고 하면, 식별자 및 준식별자에 해당하는 단어들만 비식별처리 되고, 나머지는 원문 그대로 보여지는 형태이어야 할 것이다. 따라서, 비정형 텍스트의 비식별화 결과를 다시 원래의 비정형 텍스트 형태로 원상 복귀하는 방안과 이에 따른 구현도 향후 과제로 남긴다.

References

[1] Bayardo, R. J. and Agrawal, R., “Data privacy through optimal k-anonymization,” 21st International Conference on Data Engineering (ICDE’05), 2005.

[2] El Emam, K., Dankar, F. K., Issa, R., and Jonker, E., “A Globally Optimal k-Anonymity Method for the De-Identification of Health Data,” Journal of the American Medical Informatics Association, Vol. 16, No. 5, pp. 670-682, 2009.

[3] Prasser, F. and Kohlmayer, F., “Putting Statistical Disclosure Control Into Practice: The ARX Data Anonymization Tool,” Medical Data Privacy Handbook, Springer, November 2015.

[4] Finkel, J., Grenager T., and Manning, C., “Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling,” Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics, ACL, 2005.

[5] Garfinkel, S. L., “De-Identification of Personal Information,” National Institute of Standards and Technology, 2015.

[6] Gobbel, G. T., Garvin, J., Reeves, R., Cronin, R. M., Heavirland, J., Williams, J., Weaver, A., Jayaramaraja, S., Giuse, D., Speroff, T., Brown, S. H., Xu, H., and Matheny, M. E., “Assisted annotation of medical free text using RapTAT,” Journal of the American Medical Informatics Association, Vol. 21, No. 5, pp. 833-841, 2014.

[7] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C. H., Mark,

R. G., Mietus, J. E., Moody, G. B., Peng, C. K., and Stanley, H. E., “PhysioBank, PhysioToolkit, and Physionet: Components of a New Research Resource for Complex Physiologic Signals, Circulation, Vol. 101, No. 23, pp. E215-20, 2000.

[8] Information and Privacy Commissioner of Ontario, “De-identification Guidelines for Structured Data,” Information and Privacy Commissioner of Ontario, 2016.

[9] Iyengar, V. S., “Transforming data to satisfy privacy constraints,” KDD ’02 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002.

[10] Lewis, David D, “Reuters-21578, Distribution 1.0,” UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection.

[11] Neamatullah, I., Douglass, M., Lehman, L. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., and Clifford, G. D., “Automated De-Identification of Free-Text Medical Records,” BMC Medical Informatics and Decision Making, 2008.

[12] Office for Civil Rights, “Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act(HIPAA) Privacy Rule,” U.S. Department of Health & Human Services, 2015.

[13] Park, C. W., Kim, J. W., and Kwon, H. J., “An Empirical Research on Information Privacy Risks and Policy Model in the Big data Era,” The Journal of Society for e-Business Studies, Vol. 21, No. 1, pp. 131-145, 2016.

[14] Ro, G. and Chun, J. H., “Classification and Performance Evaluation of Personal Identifiers and Quasi-identifiers for Implementing Medical Unstructured Text De-identification System,” KDBC, 2018.

[15] Sweeney, L., “Achieving k-anonymity Privacy Protection Using Generalization and Suppression,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 10, No .5, pp. 571-588, 2002.

[16] Sweeney, L., “k-anonymity: a model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 10, No. 5, pp. 557-570, 2002.

저 자 소 개

노 건 (E-mail: laylow861@gmail.com)
2015년 명지대학교 컴퓨터공학과 (학사)
2015년 ㈜프람트테크놀로지 입사
2019년 명지대학교 컴퓨터공학과 (석사)
2015년~현재 ㈜프람트테크놀로지 책임연구원
관심분야 데이터베이스, 비식별화

전종훈 (E-mail: jchun@mju.ac.kr)
1986년 University of Denver 전산과학과 (학사)
1988년 Northwestern University 컴퓨터공학과 (석사)
1992년 Northwestern University 컴퓨터공학과 (박사)
1992~1995년 University of Central Oklahoma 전산과학과 조교수
1995년~현재 명지대학교 융합소프트웨어학부 교수
2011년~현재 ㈜프람트테크놀로지 대표이사
관심분야 데이터베이스, 지능형 소프트웨어