제 출 문

미래창조과학부장관 귀하

‘개인정보 비식별 자료 생성·유통의 현장 실증 적용과제’(연구개발 기간 : 2016. 10. 18 ~ 2017. 4. 30) 과제의 최종보고서 20부를 제출합니다.

2017. 5. 15.

주관연구기관명 : SK텔레콤

대표자 : 박정호 (인)

주관 연구기관 책임자 : 김정선

참여연구원 :

SKT : 김정선, 박성준, 강미나, 곽태영, 김유경, 박건열, 박은홍, 안홍식, 이기숙, 정대인, 조태근, 한승엽

이지서티 : 김동례, 서우석, 강신곤, 윤종명, 김동호, 김은성, 김성욱, 고청천, 이주안, 이상은, 김동균, 강영준, 원병조

그리즐리 : 우호진, 최승아, 김태형, 이효복

연세대학교 : 이원석, 장영진, 이한주, JIN SHUYAN, 오현석, 용우석, 김한주

보고서 요약서

<table>
<tr><td>과제 고유 번호</td><td>2016K000334</td><td>해당 단계
연구 기간</td><td>6개월</td><td>단계구분</td><td>완료</td></tr>
<tr><td rowspan="2">연구사업명</td><td>중사업명</td><td colspan="4"></td></tr>
<tr><td>세부사업명</td><td colspan="4">미래성장동력 플래그십 프로젝트 사업</td></tr>
<tr><td rowspan="2">연구과제명</td><td>대과제명</td><td colspan="4">미래성장동력 플래그십 프로젝트</td></tr>
<tr><td>세부과제명</td><td colspan="4">개인정보 비식별 자료 생성. 유통의 현장 적용을 위한 실증</td></tr>
<tr><td rowspan="2">연구책임자</td><td rowspan="2">김정선</td><td>해당단계
참여
연구원 수</td><td>총 : 33 명
내부 : 명
외부 : 명</td><td>해당단계
연구개발비</td><td>정부: 1,000,000천원
기업: 722,000천원
계 : 1,722,000천원</td></tr>
<tr><td>총 연구기간
참여
연구원 수</td><td>총 : 33 명
내부 : 명
외부 : 명</td><td>총 연구개발비</td><td>상동</td></tr>
<tr><td>연구기관명 및
소속 부서명</td><td colspan="3">SK텔레콤 Data사업본부</td><td colspan="2">참여기업명
이지서티, 그리즐리, 연세대학교</td></tr>
<tr><td>국제공동연구</td><td colspan="3">상대국명:</td><td colspan="2">상대국 연구기관명:</td></tr>
<tr><td>위탁연구</td><td colspan="3">연구기관명:</td><td colspan="2">연구책임자:</td></tr>
<tr><td colspan="3">이동통신 데이터의 개인정보 비식별화를 통한 유통 거래 환경
구축 및 거래 실증</td><td colspan="3">보고서 면수</td></tr>
</table>

국문 요약문

연구의 목적 및 내용	본 연구에서는 기업이 보유한 다양한 원시 빅데이터에서 민감한 개인정보를 다양한 비식별화 방법을 적용하여 완전하게 제거한 비식별 빅데이터를 자유로운 유통 플랫폼을 통해서 안전하게 교환할 수 있는 새로운 체계를 실제 기업 환경에 실증적으로 적용하는 것을 목표로 함				
연구개발성과	3건의 이동통신 비식별 데이터를 유통환경에 게시하고 비식별 조치 가이드라인을 이행하여 솔루션 검증 및 재식별 안정성을 검토하였으며 이종 데이터간의 연결을 통해 빅데이터 거래 실증을 하였음				
연구개발성과의 활용계획 (기대효과)	신규 데이터 거래유통시장을 창출하고 개인 및 중소업체의 데이터 기반 서비스 및 사업 활성화에 기여				
핵심어 (5개 이내)	데이터 유통/ 플랫폼	개인정보 비식별화	비식별 조치 적정성	k-익명성	빅데이터 활용기반

〈 SUMMARY 〉

Purpose & Contents	Practical Field-applicability substantiation for creation and distribution of anonymized personal data				
Results	Publishes three set of Telco's anonymized data to distribution environment and review anonymized action & process guidelines to verify solution validation and re-identification stability Finally, demonstrate Bigdata distribution through connection between heterogeneous industry data set				
Expected Contribution	data-centric businesses and data market of small/medium size companies are newly expanded by creating a new type of data distribution markets				
Keywords	Data circulation/ platform	Anonymized Private Data	Anonymize Adequacy	K-Anonymity	Big Data

〈 목 차 〉

〈 영문 목차 〉

〈 표 목 차 〉

〈 그 림 목 차 〉

제1장 연구개발 과제의 개요

1. 연구개발 목적 및 필요성

가. 개인정보 정의 및 분류

(1) 개인정보란

> 정보통신기술이 발달하고 정보화가 촉진되면서 보호되어야 할 개인정보 유형도 다양화되는 추세임. 예를 들어 ICT 기술의 발전에 따라 수집이 가능한 위치정보가 개인정보의 영역에 포함되어 2005년에는 '위치정보의 보호 및 이용 등에 관한 법률'이 새롭게 제정됨(개인정보보호위원회 2012).

- ○ 쉽게 개인을 식별할 수 있는 정보(이름, 전화번호, 주소, 생년월일, 사진 등)
- ○ 고유식별정보(주민등록번호, 여권번호, 운전면허번호, 외국인등록번호 등)
- ○ 생체정보(지문, 홍체, DNA 정보 등)
- ○ 이외에도 사상이나 신념, 정당 가입, 건강, 성생활, 유전정보 등 다른 정보와 연결돼 개인이 확인될 수 있는 민감 정보까지를 총칭

나. 개인정보의 분류

<표 1-1> 법률상 개인정보의 정의

구 분	내용
개인정보보호법	(제2조 1) "개인정보"란 살아 있는 개인에 관한 정보로서 성명, 주민등록번호 및 영상 등을 통하여 개인을 알아볼 수 있는 정보(해당 정보만으로는 특정 개인을 알아볼 수 없더라도 다른 정보와 쉽게 결합하여 알아볼 수 있는 것을 포함한다)를 말한다.
정보통신망 이용촉진 및 정보보호 등에 관한 법률	(제2조 6) "개인정보"란 생존하는 개인에 관한 정보로서 성명, 주민등록번호 등에 의하여 특정한 개인을 알아볼 수 있는 부호, 문자, 음성, 음향 및 영상 등의 정보(해당 정보만으로는 특정 개인을 알아볼 수 없어도 다른 정보와 쉽게 결합하여 알아볼 수 있는 경우에는 그 정보를 포함한다)를 말한다.
신용정보의 이용 및 보호에 관한 법률	(제2조 1) "신용정보"란 금융거래 등 상거래에 있어서 거래 상대방의 신용을 판단할 때 필요한 다음 각 목의 정보로서 대통령령으로 정하는 정보를 말한다. 가. 특정 신용정보주체를 식별할 수 있는 정보 나. 신용정보주체의 거래내용을 판단할 수 있는 정보 다. 신용정보주체의 신용도를 판단할 수 있는 정보 라. 신용정보주체의 신용거래능력을 판단할 수 있는 정보 마. 그 밖에 가목부터 라목까지와 유사한 정보
위치정보의 보호 및 이용 등에 관한 법률	(제2조 2) "개인위치정보"라 함은 특정 개인의 위치정보(위치정보만으로는 특정개인의 위치를 알 수 없는 경우에도 다른 정보와 용이하게 결합하여 특정 개인의 위치를 알 수 있는 것을 포함한다)를 말한다.

<표 1-2> 개인정보보호 가이드라인 제·개정 현황

제 목	법정근거	제정(개정)월	비고
공공정보 개방·공유에 따른 개인정보보호 지침	-	2013.9	-
금융분야 개인정보보호 가이드라인	개인정보 보호법 제12조	2013.12	금융위원회, 금융감독원 공동
약국 개인정보보호 가이드라인		2013.12	보건복지부 공동
사회복지시설 개인정보보호 가이드라인		2013.12	보건복지부 공동
의료기관 개인정보보호 가이드라인		2013.12	보건복지부 공동

다. 개인정보 비식별화 현황 및 문제점

현재 개인정보 관련 법제도의 강한 규제로 기업현장에서는 정보사용에 있어 매우 소극적이며, 개인정보 논란을 피하기 위해 통계정보 제공 수준의 정보이용만 가능하여 정보의 가치가 떨어지고 사업적으로 어려움이 크다. 기업에서는 법적 검토를 통해 데이터 활용 문제가 없더라도 고객 관련정보의 개방을 할 수 없을 뿐더러, 현재 개인정보 수집, 관리 검증 체계 및 지원이 미흡하여 개인정보 관리 및 데이터 서비스에 어려움이 되고 있다.

○ 개인정보의 비식별화가 충분하지 않은 상태에서 광고나 마케팅 등에 무분별한 개인정보 활용으로 인한 문제 발생

- 타켓 광고: 비식별 개인정보를 이용한 신종 마케팅 기법, 리타게팅 광고*가 성행하면서 헌법상 기본권인 개인정보자기결정권**이 침해되고 있다는 주장이 제기되며 개인정보 자기결정권 침해 논란

* 리타게팅 광고 : 인터넷 사용자들의 웹페이지 방문, 검색 기록 등 비식별 개인정보를 바탕으로 웹사이트 방문자에게 해당 콘텐츠를 보여주는 마케팅 기법

** 개인정보자기결정권 : 자신에 관한 정보를 보호받기 위하여 자신에 관한 정보를 자율적으로 결정하고 관리할 수 있는 권리

(1) 각종 개인정보 유출 사례

○ KT 1,170만건('14.3월), 카드3사 8,700만건('14.1월), SK 컴즈 3,500만명('11.7월), 옥션 1,860만명('08.2월), 넥슨 1,320만명 유출('11.11월), GS칼텍스 1,150만명('08.9월) 등

○ 현대캐피탈 서버를 해킹하여 개인정보를 유출한 해커 기소 (2013.1)

○ 코웨이 고객 198만 명의 개인정보 유출 확인 (2013년 2월)

○ 개인정보 유출 피해자들, 주민등록번호 변경 헌법소원 제기 (2013년 2월)

○ 북한 공작원, 해커들과 공모해 개인정보를 해킹 및 거래해 온 국내 불법 사이트 운영자 구속기소(2013년 4월)

○ 정부기관 홈페이지 등 해킹(2013년 6월)

(2) 각종 개인정보의 불법적인 유통

○ 대출·통신·텔레마케팅 등 중심으로 금전적 이득을 위한 개인정보 불법유통 시장이 광범위하게 형성된 것으로 추정

○ 전화번호 정보는 건당 10원 ~ 50원, 대출기록이 포함된 개인정보는 건당 5000원 ~ 2만원에 판매('14.01.08, 서울신문)

○ 개인정보 불법유통 게시물 검색건수(인터넷진흥원) : '13년 18,994건, '14년 6월 20,192건

라. 개인정보 비식별화의 필요성

정부 및 기업에서는 빅데이터를 활용한 비즈니스 관심이 증가하고, 수집정보를 활용한 가치 창출에 대한 기대치가 증대하고 있다. 하지만, 국내 빅데이터 시장은 개인정보보호법 제도 하에서 발생될 법적 분쟁을 염려하여 외부데이터 활용에 소극적인 상황이다. 이에 개인정보의 안전한 가공 및 변환을 통해 보호와 활용의 균형을 모색할 수 있는 절차적·기술적 개선방안 도출이 절실하다. 특히, 정부 차원에서도 성공적인 정부3.0 추진을 위해 개인정보보호에 대한 현 수준을 진단하고 문제점을 해결하기 위한 전략수립의 일환으로 실제 개인정보 데이터를 바탕으로 비식별화 및 유통활성화 방안이 필요하다.

(1) 데이터 기반 정책 활성화

○ 데이터 기반의 정책과정과 의사결정은 전략적 국가과제 수립에 대한 위험성을 감소시키는 효과가 있으나 국민 개개인의 사생활이 침해될 수 있기 때문에 개인정보보호가 선행되어야 함

○ 데이터 기반의 정책수립과 의사결정으로 정책의 효과성 제고하는 것이 전세계적 추세

○ 데이터 기반의 정책과정과 의사결정은 전략적 국가과제 수립에 대한 위험성을 감소 시켜주는 효과

○ 다만, 빅데이터의 선순환적 기능과 반하여 국민 개개인의 사생활이 침해 될 수 있는 가능성 존재

(2) 개인정보 비식별화 영역의 확장

○ 개인신상정보 뿐만 아니라 개인 생활로그정보로 비식별화 영역을 확장함으로써 분석의 스펙트럼을 확장시키고 다양하고 분석결과의 도출 및 적용 필요

○ 개인 신상정보에 이은 개인 생활로그정보의 비식별화 분석 가능 빅데이터의 영역 확장

○ 빅데이터 분석 영역의 확장으로 분석 스펙트럼이 확장되며, 공공 및 여러 산업분야에서 다양하고 깊이있는 분석 결과의 도출 및 적용이 필요함

(3) 정책에 부합되지 않는 빅데이터 활용

○ 공공 데이터, 산업 데이터 등 빅데이터 활용에 대한 정부 및 산업의 수요에 따라 2013년 9월 안전행정부 '공공정보 개방, 공유에 따른 개인정보 보호 지침', 2014년 5월 미래창조과학부 '빅데이터 활용을 위한 개인정보 비식별화 사례집', 2014년 12월 방송통신위원회

‘빅데이터 개인보호 가이드라인‘, 2015년 5월 한국정보화진흥원 ’개인정보 비식별화 기술 활용 안내서‘ 발간 등 계속적인 노력에도 빅데이터 활용을 위한 현실적인 부분에 부합되지 않음

○ 2016년 6월 미래창조과학부, 국무조정실, 행정자치부, 방송통신위원회, 금융위원회, 미래창조과학부, 보건복지부가 합동으로 ‘개인정보 비식별 조치 가이드’의 발간을 통해 빅데이터 유통 문제를 적극적으로 해결하기 위해 노력을 하고 있음

○ 정부관계부처의 노력에도 아직 비식별화의 기술적 문제가 남아 있으며, 본 과제의 실증을 통하여 빅데이터 유통상의 발생할 수 있는 문제를 해결하고자 함

마. 빅데이터 유통 활성화의 필요성

(1) 빅데이터 유통 활성화

○ 빅데이터 유통 활성화를 위해서는 개인정보 비식별화가 반드시 전제되어야 함

○ 비식별 가이드와 기술을 적용하여 개인정보가 안전하게 처리된 비식별 빅데이터를 안전하게 유통함으로써 빅데이터 산업을 활성화시킬 수 있음

○ 유통 플랫폼을 기반으로 빅데이터 분석을 통하여 유의미한 정보를 추출하기 위해서는 개인 신상 데이터 뿐 아니라 로그성 개인 생활데이터가 연계되어야 하며, 빅데이터간 연계 또한 필요함

(2) 데이터 거래 유통의 필요성

○ 데이터의 증대와 활용성이 중시되면서 데이터를 중심으로 하는 ‘데이터 생태계’가 빠르게 확산 중

○ ‘빅데이터와 오픈 데이터’ 등 데이터의 시대를 맞아 ICT의 주도권이 ‘인프라’에서 ‘데이터’로 이동하면서 데이터 생태계가 형성

○ 데이터의 활용이 중시되면서, 데이터의 생산과 관련 없는 기업·기관이 데이터를 수집·가공· 유통, 소비자가 될 수 있는 구조가 형성

○ 생태계에서 생산과 소비를 이어주는 ‘유통’이 강조되기 시작

(3) 데이터 거래 기반조성의 필요성

○ 데이터 유통을 위한 모델 수립과 데이터 유통 플랫폼을 확대하여 거래가 활성화 될 수 있도록 기반 조성 추진이 필요

○ 유통 가능한 데이터를 상황에 맞게, 어떠한 방식으로 유통할 것인지에 대한 내용이 필요

○ 스타트업들은 비용 부족으로 데이터 이용이 어려우므로 유통 플랫폼을 개발하여 데이터 유통 활성화가 필요

○ 현재 데이터 유통을 위해서는 개인정보에 관한 식별정보 유통은 불가능한 상황

(4) 데이터 비식별화 기술적·제도적 기반 조성의 필요성

○ 식별정보 등은 마스킹, 암호화 등 익명성을 전제로 하고 유통할 수 있는 비즈니스 적 모색 필요

○ 데이터 산업의 활성화를 위해 개인정보 비식별화 기술을 검증하고 이를 통해 재식별의 리스크 없는 데이터 마켓 플레이스를 구축하여 다양한 데이터들의 거래 실증 필요

2. 연구개발 범위

가. 원본 빅데이터 수집 및 제공(SK텔레콤)

(1) SKT의 원본 데이터 가공

○ SKT에서 보유 중인 가입자 정보와 통화요금 수납정보, 위치 정보를 중심으로 가입자 연체 및 신용도 유관 데이터, 장애우 및 외국인 가입자 체류지/체류시간 데이터 3종 개발, 가공하여 제공(10-15만건 Sample 데이터 제공)

○ 제3자 정보제공 동의자를 대상으로 정보 가공하며, 비식별 처리를 위한 협약기관간 위탁계약체결 및 개인정보처리방침 고지 후 진행

○ SKT 가입자 중 20대 청년 이용자의 요금제, 수미납 연체이력

○ SKT 가입자 정보를 이용한 외국인의 관공지 체류장소.시간 정보를 가공

○ SKT 가입자의 복지할인 유형 데이터를 활용한 사회 약자의 거소지 정보를 가공

나. 비식별화 알고리즘 수행 및 검증

(1) KLT 프라이버시 모델 실증(이지서티)

○ KLT 비식별화는 개인의 식별할 수 있는 식별자를 제거하여 개인을 식별할 수 없도록 조치하는 것을 비식별화라 함

○ 프라이버시 보존을 위한 비식별 조치 검증 기술 적용

○ 비식별 조치된 정보 집합물의 k-익명성 검증 기술 적용

○ k-익명성 검증을 위해서 각 레코드에 맵을 생성한 후 전체 데이터 집합을 검사하여 맵을 체크

○ 맵이 동일하면 해당 맵을 카운트(Count)하고 최종 카운트(Count)값을 검사하여 맵이 설정된 K값을 만족하는지 여부를 검사하는 방식을 통해 k-익명성 검증

(2) MAS 비식별화 알고리즘 실증(그리즐리)

○ 기존의 KLT 비식별화 기법은 재식별 불가성 한계가 있으며, 데이터 간 연계를 지원하지 않는다는 한계점이 존재

○ 이러한 단점을 해결하고 보다 안전한 비식별화 변환 기법을 적용하여 데이터 유통 및 연계 활용을 지원하기 위하여 새로운 기술인 MAS(추상동기화) 기법의 비식별 조치 검증 기술 적용

○ MAS 기법에 의한 추상화 비식별 데이터의 생성을 통한 비식별 불가성 검증 및 동기화 기법에 의한 이종의 데이터 결합 가능성 검증

다. 비식별·연계 데이터 검토 및 검증(연세대)

(1) 비식별화 기법 별 적정성 평가 프로세스 적용

○ 개인정보 비식별 조치 가이드라인의 발표에 따라 전문적인 비식별 조치의 검증과 그 적정성을 평가하여 개인정보 보호 법률의 이행의 실증 필요

○ 비식별 조치 된 데이터의 연계 및 데이터 활용 시 재식별 가능성을 고려하여 재식별 불가성을 검증하여 그 안전성을 보장

○ 서로 다른 비식별 조치 된 데이터의 연계 시 그 정확성을 검증하며 데이터 활용 및 분석 결과의 정확성과 효율성을 확인

라. 비식별 데이터 유통 플랫폼 구축 및 유통 실증(SK텔레콤)

(1) 비식별 빅데이터 유통 플랫폼 구축

○ SK텔레콤이 운영 관리중인 인프라 환경을 바탕으로 유통 플래폼 개발

○ SK텔레콤 가입자 서비스 이용 빅데이터에 대한 국내 비식별화 기술 적용으로 재식별 가능성 없는 안전한 비식별 데이터 유통 플랫폼 구축

(2) 비식별 조치 데이터 유통 실증

○ 개인정보의 재식별이 불가한 비식별화 기술의 적정성 평가를 통한 안정성 검증

○ 이종 산업의 데이터와의 결합 및 유통 실증을 통한 비식별 데이터의 활용 가능성 검증

제2장 국내외 기술 개발 현황

1. 기술현황

가. 빅데이터 관련 국내 공공/기업 활용

(1) KT 빅데이터 사업 확산

○ 빅데이터 전담인력을 상한선 없이 충원할 계획이며, 내·외부를 통틀어 진행 중인 프로젝트 20개를 40여개로 늘릴 예정

○ 지난 1년간 개발한 빅데이터 융합 모델이 성공적이라는 평가를 받음. 정부 규제는 문제가 되지 않고 오히려 융합 모델을 만들 예정

(2) 필립스 전자 빅데이터 활용 사례

○ 2009년 이유식 제조기는 재고가 없어서 못 팔 만큼 인기를 모았지만, 2011년 제품 판매량이 급감

○ 필립스 내부에선 유사 저가 제품 탓이라 생각하여 가격인하를 전략으로 내세우려 했지만, 데이터(1억 4000만건 육아 블로그, 사이트) 분석 결과, 배달이유식 영향으로 밝혀져, 가격변동, 기능 업그레이드가 아닌 광고 메시지 변경으로 6개월 만에 시장 1위 탈환

(3) 교보생명 빅데이터 활용 사례

○ 기계약 고객들에게 피드백을 받아 예측모델을 만든 뒤 설계사후보자들에게 적용해 활동가능성과 우수설계사가 될 가능성을 점수화하여 각 점포단위로 배분하여 리쿠르팅 활동을 수행하도록 함. 또한 약관대출 잠재고객 도출에도 빅데이터 사용 중

(4) 관세청, 품목분류 빅데이터 민간에 무료 개방

○ 세관에서 관세 등을 물릴 때 적용하는 품목분류(Harmonized System, HS) 데이터 개방

○ ▲48개국 2014년 관세율표 ▲HS가이드 ▲HS국제분쟁 해결 사례 ▲HS관련 법령정보 등 81만여 건의 엑셀, 한글파일 제공(해당국가 원문·영문·국문으로 번역)

(5) 식품의약품안전처, 빅데이터 활용한 통계정보 시각화 사이트 오픈

○ 식품의약품, 화장품, 의료기기 등 분야별 생산 실적 정보를 시도 및 시군구 단위의 통계지도, 차트 형태로 제공하는 사이트 오픈

○ 최근 5년간의 변화를 비교·분석할 수 있는 차트를 제공하며 차트 이미지 및 원본 자료 다운로드 가능

(6) LX공사, 사고 예방용 급경사 알림 서비스

○ 국도경사도와 지질도, 지형도 등 관련 정보와 기상정보, 유동인구 등을 융·복합해 급경사지 위험도를 산정하는 빅데이터 모델을 개발하는 사업 진행할 예정

○ 서비스가 상용화되면 폭우, 폭설 등 기상재해 발생 시 사고예상지역에 대한 선제적인 대국민 정보 제공 가능할 것으로 기대

나. 개인정보 비식별화 기술 현황

(1) 가명처리

○ 개인식별이 가능한 데이터에 대하여 직접적으로 식별할 수 없는 다른 값으로 대체하는 기법

○ 휴리스틱 익명화(데이터의 유용성이 떨어지고 활용가능한 대체 변수의 한계가 있음)

○ 암호화(양방향 및 단방향 암호화)

○ 스와핑(사전에 정해진 외부 값으로 대체)

(2) 총계처리

○ 개인정보에 대하여 통계값(전체 혹은 부분을 적용하여 특정 개인을 판단할 수 없도록 함)

○ 총계처리(단체의 속성이 개인정보를 그대로 대변하는 경우가 발생 가능)

○ 부분집계(오차범위가 큰 항목, 속성 값을 변환 ex.-40대 소득)

○ 라운딩(올림, 내림처리)

○ 데이터 재배열(나이, 소득 등 특정 속성을 개인별 교환)

(3) 데이터 삭제

○ 개인정보 식별이 가능한 특정 데이터 값을 삭제 처리

○ 속성값 삭제(개인식별 항목을 단순제거)

○ 속성값 부분삭제(민감한 속성 값 일부 삭제)

○ 데이터 행 삭제(해당 정보를 가진 개인의 내용을 전체 제거)

○ 준식별자 제거를 통한 단순 익명화(잠재적 개인 식별이 가능한 준식별자 모두를 제거)

(4) 데이터 범주화

○ 단일 식별 정보를 해당 그룹의 대표값으로 변환하거나 구간값으로 변환하여 고유정보 추적 및 식별 방지

○ 범주화(명확한 값을 숨기기 위해 데이터 평균또는 범주값으로 변환)

○ 랜덤올림(총계처리의 라운딩과 같은 방법)

○ 범위방법(임의의 수 범위로 구간 값 설정)

○ 제어 올림방법(랜덤 올림 방식에서 행과 열의 합에 대한 일치 해결)

(5) 데이터 마스킹

○ 개인식별 정보에 대하여 전체 혹은 부분에 대한 대체값을 공백 혹은 *, 노이즈 등으로 변환

○ 임의 잡음 추가(민감정보에 임의의 숫자 등 잡음을 추가)

○ 공백과 대체 방법(비식별 대상 데이터 선택 후 선택된 항목 공백 또는 특수문자 등으로 대체)

다. 프라이버시 모델

(1) k-익명성 모델

○ 02년 sweeny에 의해 제안된 모델로 가장 대중적으로 사용되는 데이터 공유 모델

○ 데이터의 속성들에 대하여 준식별자 속성을 선정하고 이에 대한 준식별자 속성값의 조합이 동일한 레코드의 집합을 동질집합이라고 명명

○ 동질집합의 레코드 개수는 항상 k개 이상 존재함을 보장하는 프라이버시 모델

○ 준식별자 속성값 조합을 근거준식별자에 대해 generalization과 suppression을 수행하여 모델을 달성함

- 준식별자 : 속성들 중 식별자는 아니지만 일반적으로 알려진 개인을 묘사할 수 있는 속성들로 일부 조합을 통해 개인을 식별할 수 있는 가능성이 높은 속성들을 의미함 (예:성별, 나이, 주소)
- generalization과 suppression: 일반적으로 속성값을 트리로 표현하여 상대적으로 자식 노드에 위치한 속성값을 부모 노드의 값으로 대체하는 비식별화 기법

(2) l-다양성 모델

○ 2007년 Ashwin Machanavajjhala에 의해 제안된 모델로 k-익명성 모델의 취약점을 보완한 모델

○ 특정 동질집합의 민감속성 조합이 모두 동일할 경우 해당 동질집합의 민감속성의 값이 바로 드러나는 문제를 해결하기 위한 프라이버시 모델

○ 데이터의 속성들중 준식별자와 함께 민감속성을 선정하고, 동질집합 내에서 민감속성의 값이 k개 이상 보존됨을 보장하는 프라이버시 모델

○ 준식별자, 민감속성의 속성값 조합을 대상으로 위험을 판별하고 준식별자에 대해 generalization과 suppression을 수행하여 모델을 달성함

- 민감속성: 전체 속성들 중 준식별자를 제외한 속성들 중에서 일반적으로 유출되어서는 안된다고 여겨지거나 도메인에 따라서 지정되는 반드시 보호되어야할 속성들의 집합을 의미함

(3) t-근접성 모델

○ 2007년 N Li에 의해 제안된 모델로 k-익명성,l-다양성 모델의 취약점을 보완한 모델

○ 특정 동질집합의 민감속성이 l-다양성을 만족하더라도, 해당 동질집합의 레코드 분포가 전체 도메인과 대비하여 특이적이지 않음을 보장하는 모델

○ 준식별자, 민감속성의 속성값 조합을 대상으로 위험을 판별하고 준식별자에 대해 generalization과 suppression을 수행하여 모델을 달성함

(4) 마이크로집계(microaggregation) 모델

○ 전체 레코드 세트를 작은 단위의 레코드 세트로 나누고 각 속성값에 대한 집계처리를 통해 개인정보를 보호하는 프라이버시 모델

○ 일반적으로 나누어진 집합의 최소 레코드 개수 k를 지정하여 k-익명성과 유사한 프라이버시 보호를 수행함

○ 다양한 구현 방법에 따라서 준식별자, 민감속성 등에 대한 속성 구분을 크게 짓지 않는 경우가 존재

(5) 차분 프라이버시(Differential privacy) 모델

○ 06년 Dwork에 의해 제안된 모델로 통계적 정보 유출에 대한 연구의 연장선에서 제시된 프라이버시 모델

○ 특정 개인의 데이터가 공개된 데이터 안에 존재하는지를 확률적으로 보장하는 기법으로 공격자의 사전 지식에 영향을 받지 않으면서 개인정보를 보호 함

○ 지정된 속성에 대해 랜덤한 노이즈를 추가하는 방법으로 비식별화를 수행하되, 노이즈의 범위가 분석결과에 대한 민감성과 개인정보 유출에 대한 위험성을 기반으로 지정될 수 있도록 파라미터화 되어있음

2. 시장현황

가. 국내 비식별화 기술 시장 현황

(1) 비식별 기술 시장 현황

○ 국내의 제품이나 솔루션은 대부분 비식별화 기술 중에서도 암호화 기술에 치중

○ 최근들어 데이터 마스킹 및 토큰화 기능을 제공하는 제품이 상용화 되어 출시되기 시작

(2) 국내 시장 규모

<표 2-1> 국내 비식별 기술 시장 규모

(단위:억원)

	현재년도	개발 종료 후 1년	개발종료 후 3년
시장규모(한국)	2,600	6,764	11,036

(3) 국내·외 주요 수요처 현황

<표 2-2> 국내·외 주요 수요처 현황

수요처	국명	관련제품
공공기관	전 세계	정부 각 기관에 등록된 개인정보 관리서비스
기업	전 세계	개인정보 보호 및 해당 데이터 분석
연구단체	전 세계	해당 연구 분야의 연구에 활용

나. 국내 데이터 유통 시장 현황

(1) 데이터 유통 시장 현황 및 난제

○ 정부위주의 공공데이터 오픈(data.go.kr) 및 데이터 거래환경 조성을 위한 노력(DB Store)은 지속되고 있으나 개인정보 침해에 대한 우려와 리스크가 상존하므로 개별정보에 대한 거래 환경 조성에 어려움이 있음

(2) 전문기관을 통한 비식별 조치 활성화 미흡

○ 2017년 2월까지 달까지 각 기관 별 전문가 추천 건수와 결합 건수는 한국신용정보원(6건, 3건) 금융보안원(4건, 3건) 한국정보화진흥원(5건, 2건) 한국인터넷진흥원(1건, 0건) 사회보

장정보원(1건, 0건) ▲정부통합전산센터(1건, 0건) ▲한국교육학술정보원(0건, 0건) 으로 활성화 및 정부의 기대대비 미흡한 수준임

3. 경쟁기관 현황

가. 데이터 프라이버시 관련 소프트웨어를 개발하는 기업 현황

(1) IBM

○ IBM Infosphere Optim software를 개발하여 데이터 프라이버시를 위한 시스템 개발

(2) Privacy Analytics

○ 2007년 설립되어 의료 데이터 익명화를 위한 PARAT software 개발

나. 연구 목적으로 데이터 프라이버시 관련 시스템을 개발하는 연구기관 현황

(1) TU München

○ k-익명성, l-다양성, t-근접성을 적용시킨 오픈소스 ARX-Powerful Data Anonymization Tool을 배포

(2) UT Dallas Data Security and Privacy Lab

○ Mondrian, Incognito를 고려한 k-익명성, l-다양성, t-근접성 그리고 Anatomy를 적용시킨 UTD Anonymization ToolBox 개발

다. 정보손실 방지를 위한 통계적 정확성 유지 비식별화 기술

(1) 국내의 비식별화 제품이나 솔루션은 대부분 암호화, 마스킹 기술에 치중되어 있으며, 삭제, 총계처리 등의 비식별화 변환 기술을 손쉽게 적용할 수 있는 제품 및 솔루션은 부족

○ 국내 기업 중 펜타시큐리티, 이글로벌, 이사인, 소프트포럼, 신시웨이와 같은 벤더에서 DB 암호화 솔루션을 제공

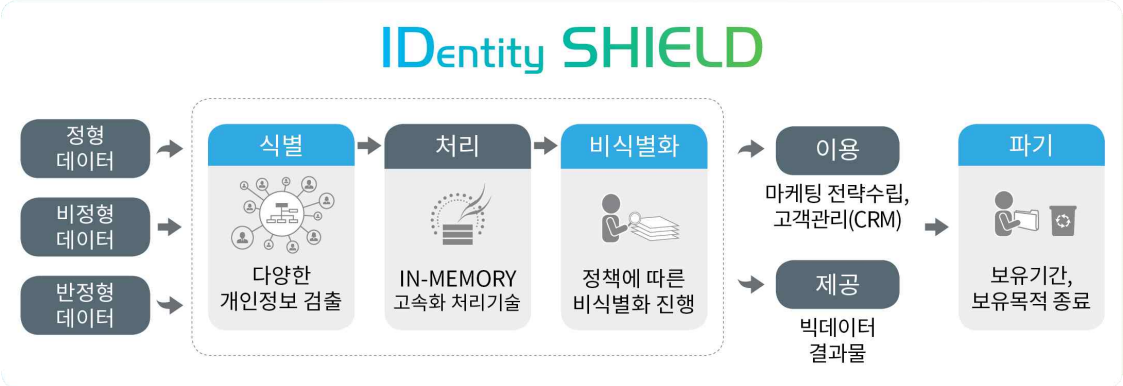
○ 웨어벨리, 소만사, 코아인포메이션 등의 벤더들이 데이터 마스킹 기술이 적용된 제품을 제공하고 있으나 실시간 비식별화 변환 기술은 제공되지 않음

라. 국내·외 상용 비식별화 솔루션

(1) 이지서티 - 아이덴터티 쉴드(IDENTITY SHIELD)

○ 아이덴티티 쉴드는 국내 비식별화 시장을 리딩하는 제품으로 비식별화 전문기관인 정부통합전산센터에 구축하여 중앙부처 및 지자체 비식별화 서비스를 구축하였으며, 서울시 캠퍼스에 비식별화 솔루션을 구축하여 운영하고 있다. 전문기관에서는 TTA(한국정보통신기술협회)를 통한 BMT를 진행하여 아이덴티티 쉴드가 유일하게 검증받은 제품 임

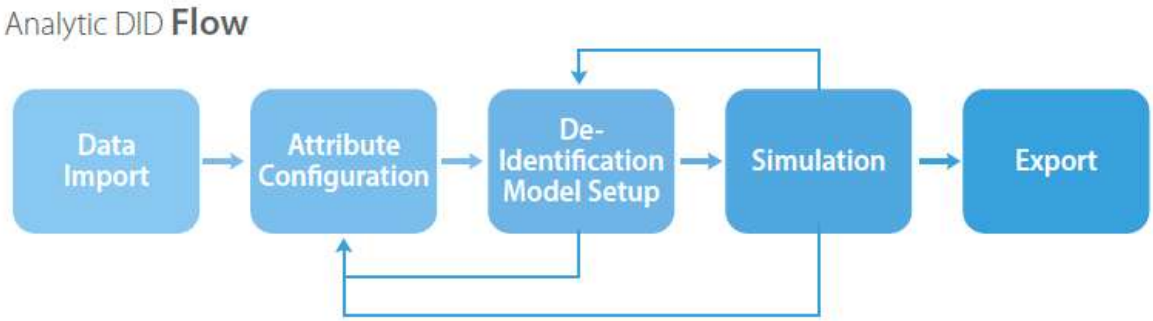
○ 아이덴티티 쉴드는 개인정보 비식별조치 가이드라인에 따라 17가지 고전적방식과 KLT프라이버시 모델을 제공하고 있음. 17가지 고전적방식은 빅데이터 전문 분석가들이 조건별 수동으로 비식별 조치를 수행함으로써 빅데이터 활용을 지원하며, KLT 프라이버시모델 적용은 자동화된 비식별조치를 수행함으로써 빅데이터 전문가가 아니더라도 편리하게 비식별조치를 수행할 수 있도록 지원함.



[그림 2-1] 이지서티 - IDENTITY SHIELD

(2) 파수닷컴 - 애널리틱 디아이디 (Analytic DID)

○ 데이터 내의 개인정보를 KLT 프라이버시모델을 통해 비식별조치 하는 솔루션이며 데이터의 유형 및 특징, 분석목적에 따른 분석을 위한 시각화를 지원



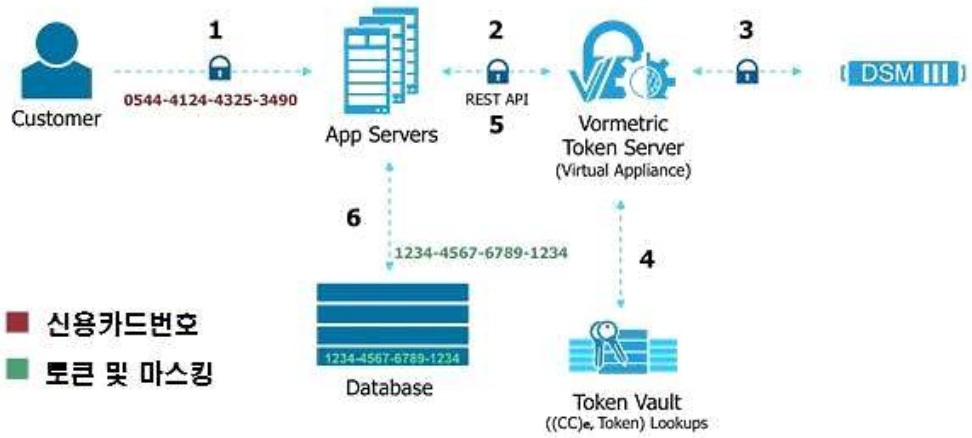
[그림 2-2] 파수닷컴 - Analytic DID

(3) 팬타시스템 - DataEye PIDI

DataEye PIDI (Privacy Information De-identification) 솔루션은 개인정보 비식별 조치 가이드라인을 준수하여 식별자 암호화 등 17가지 비식별 조치 기술을 지원하며, k-익명성, l-다양성, t-근접성의 개인정보 보호 모델을 통한 비식별 조치 적정성 평가를 수행할 수 있는 개인정보 비식별 조치 솔루션임. 데이터 변환 및 이관처리를 위한 ETL 솔루션이 내장되어 있으며 인메모리 DB를 활용한 데이터 처리하며 대용량 데이터의 비식별 조치를 위한 Incognito 알고리즘 적용한 k-anonymity(k-익명성), l-diversity (l-다양성), t-closeness (t-근접성)의 적정성평가 모델 지원함

(4) 보메트릭 코리아 - 보메트릭 토큰 서버

○ 어플리케이션 단계에서의 토큰화 기술을 통해 데이터를 가명 처리 및 마스킹 처리를 하는 기술
○ 동적인 데이터 마스킹 및 속성 유지 토큰화 기술을 이용하여 데이터의 형식을 보존하여 성능의 영향없이 민감한 정보를 보호함.



[그림 2-3] 보메트릭 - 토큰서버

4. 지식재산권 현황

가. 지식재산권 현황

○ 빅데이터에서 세션별 패킷 수집 기반 로그 생서 방법 및 장치(2016.08.19 출원, 주식회사 이지서티)

○ 개인정보 비식별화 전송장치 및 전송방법 (2015.01.19 출원, 경희대학교 산학협력단)

○ 정형 및 비정형 데이터를 포함하는 빅데이터에서의 개인정보 익명화 관리 시스템 (2014.09.04 출원, 주식회사 바넷정보기술)

○ k-익명성을 제공하는 정보 보호 방법 및 장치 (2010.12.07 출원, 서울대학교 산학협력단)

○ 정보의 익명화를 위한 데이터 저장 시스템 및 방법 (2005.12.20. 출원, 주식회사 유케어소프트)

나. 신청기관 지식재산권 현황

○ 개량된 k-익명성 모델이용 데이터셋 비식별화 방법 및 장치 (2017.04.27 출원, 주식회사 이지서티)

○ 비식별화 데이터셋 결합용 키 생성 방법 및 장치 (2017.04.27 출원, 주식회사 이지서티)

○ 빅데이터의 비식별화 처리 방법 (2016.06.09 출원, 주식회사 그리즐리)

5. 표준화 현황

가. 본 기술/제품과 직접적으로 관련 있는 국내 표준화 현황

(1) 개인정보 접근 비정상 행위 패턴, 부적절 이용 탐지 기술
(이상금융거래 탐지/TTAK.KO-12.0178)

○ 전자금융 이용자의 금융이용 환경, 금융 거래 패턴, 거래 사전 행위 분석을 통해 이상 금융 거래를 정의하고 있으며, 이상 금융 거래의 탐지를 위한 요구 사항들과 금융이용 환경, 금융 거래 패턴, 거래 사전 행위를 통한 이상 금융 거래 탐지 및 대응방법을 제시

○ 거래 패턴 기반 탐지 및 방법의 예 : 금융거래패턴에 의한 탐지는 금융거래 이용자의 과거 거래패턴을 기반으로 이상금융거래를 탐지하여, 이상금융거래 탐지 여부에 따라 금융거래 패턴 정보는 지속적으로 재설정 되어야 함

(2) 개인정보 접근 비정상 행위 패턴, 부적절 이용 탐지 기술
(침입 탐지 시스템 기능 패키지/TTAS.KO-12.0026)

○ 침입 탐지 시스템의 보안 요구사항을 서술하는 보호 프로파일이나 보안목표 명세서를 작성할 때 참고할 수 있는 기능 패키지

제3장 연구 수행 내용 및 성과

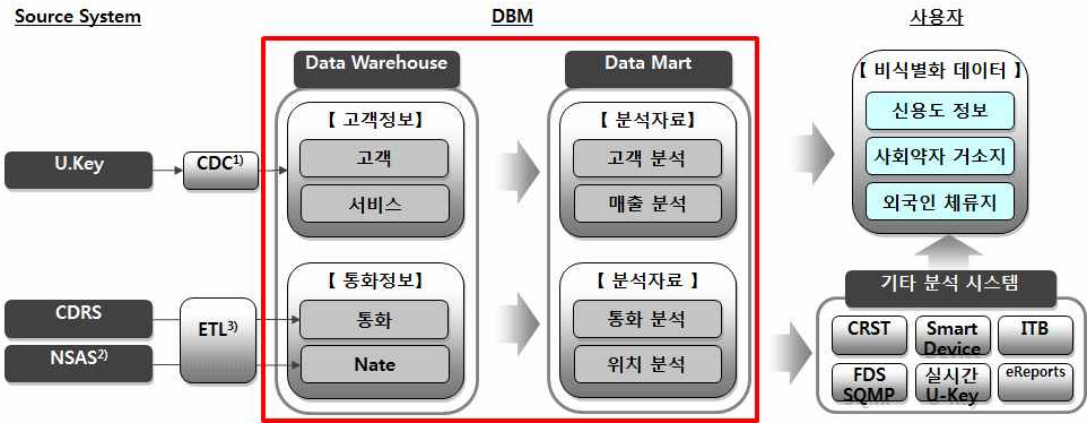
1. 데이터 생성 가공

가. 데이터 생성 가공

(1) DBM(Databased Business Management) 시스템 개요

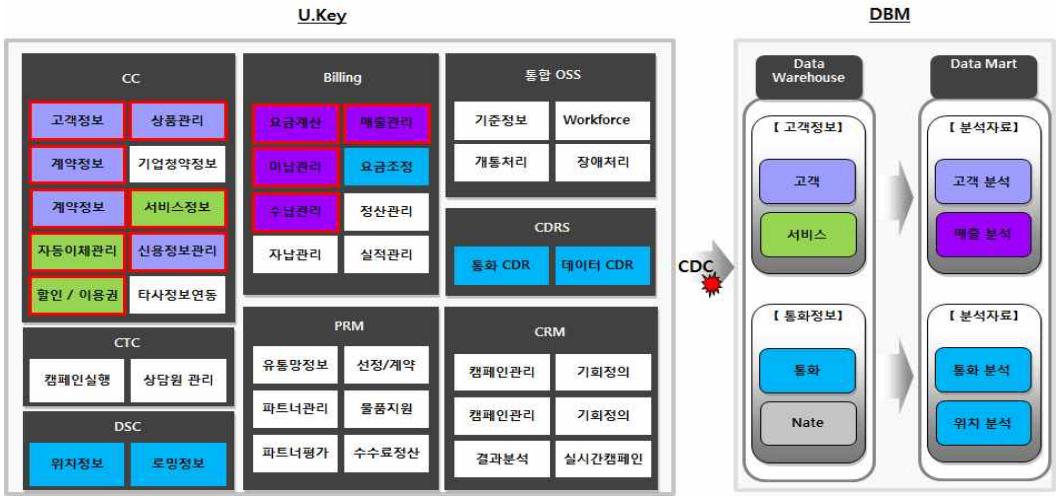
SKT 고객 정보 기반으로 축적된 DW(Data Warehouse) 데이터로 다양한 분석을 통해 마케팅 및 경영활동 전반에서 필요로 하는 분석/통계 정보를 제공하는 시스템으로, 대용량 데이터와 다차원 분석환경을 제공

비식별용 신용도 정보, 사회약자 거소지 정보, 외국인 체류지 정보를 분석/ 추출시 DBM 시스템을 이용



(2) 비식별용 데이터 기초 정보 분석 및 적재(U.key 시스템)

SKT 고객정보시스템인 U.Key로 부터 DBM으로 주기적으로 데이터를 적재하고 있으며, 비식별용 기초 데이터 분석 및 추출하기 위해서 고객정보/통화정보/청구정보/미납정보/위치정보 등을 추가적으로 필요한 정보는 신규로 데이터를 가져와서 DBM System에 적재해야 함

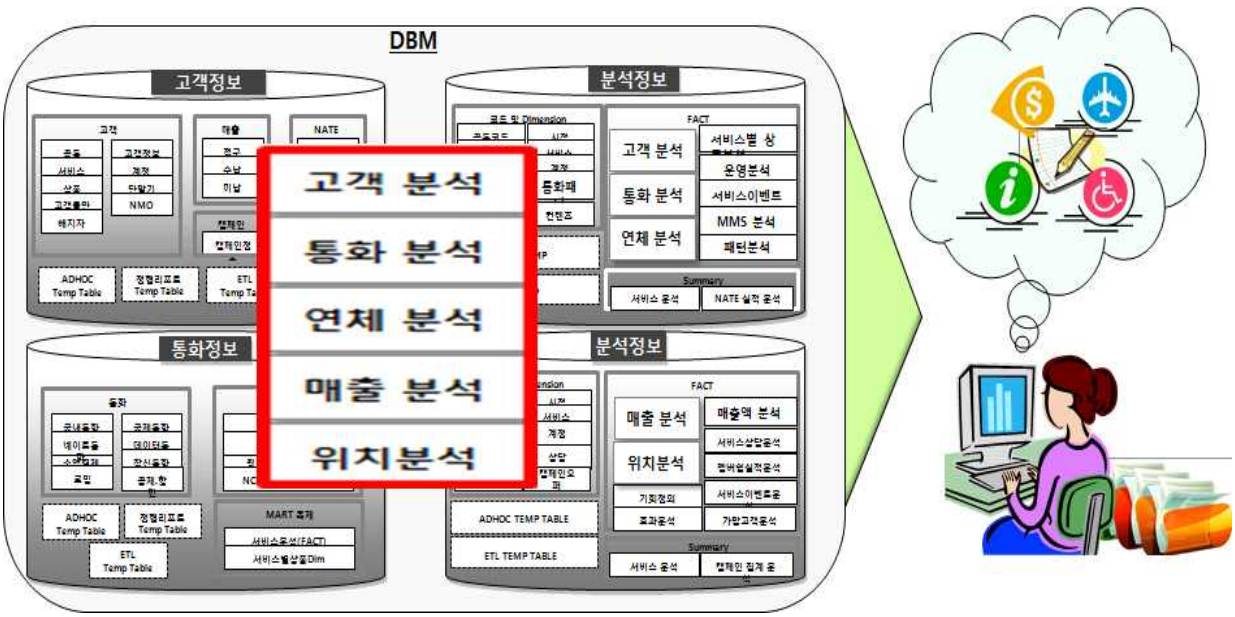


(3) 비식별용 데이터 분석/설계(DBM)

DBM에 적재된 고객정보, 통화정보, 수/미납 정보 등을 통해 분석된 다양한 종류의 DBM 자료를 분석. DBM의 분석된 자료는 비식별용 데이터을 추출하기 위한 설계 자료가 됨

고객 분류	데이터	제공 서비스	
상품 분석	▪고객의 사용 요금제, 사용 단말기 등의 내용 분석	마케팅 OLAP	A D H O C
위치 분석	▪사용자의 통화 위치 및 거소지 정보 분석		
고객 분석	▪고객 정보 분석(연령별, 지역별, 요금제별 등)	캠페인 OLAP	
통화 분석	▪고객별 통화량 분석(월별, 일별, 시간대별 등)		
매출 분석	▪과거 trend 파악 (36개월 통계를 통한 trend파악 예:월별/년별 가입자 변동)	BI portal	
NATE분석	▪scoring(통계 로직을 적용하여 스코어링:고객등급,번호이동 예측 점수등)		
수/미납분석	▪고객의 수/미납 정보 분석 ▪연체 정보 분석	마이닝	
상담분석	▪고객센터를 통해 접수/처리된 내용 분석		

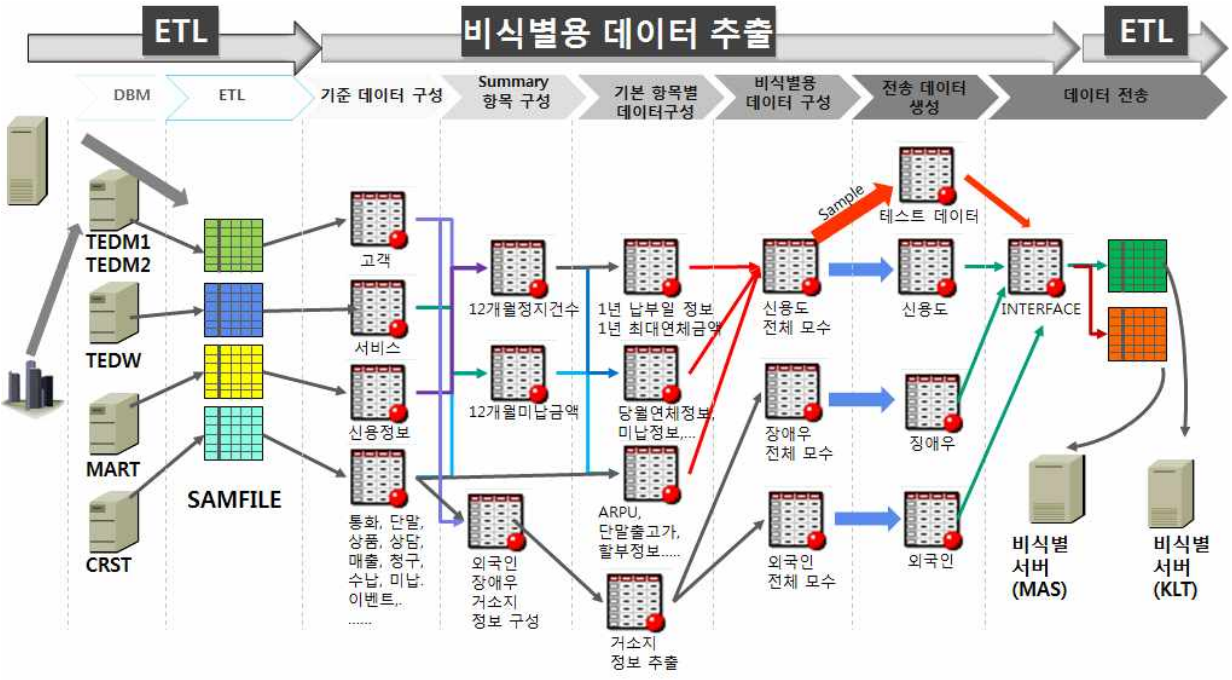
(4) DBM System의 Table 분석



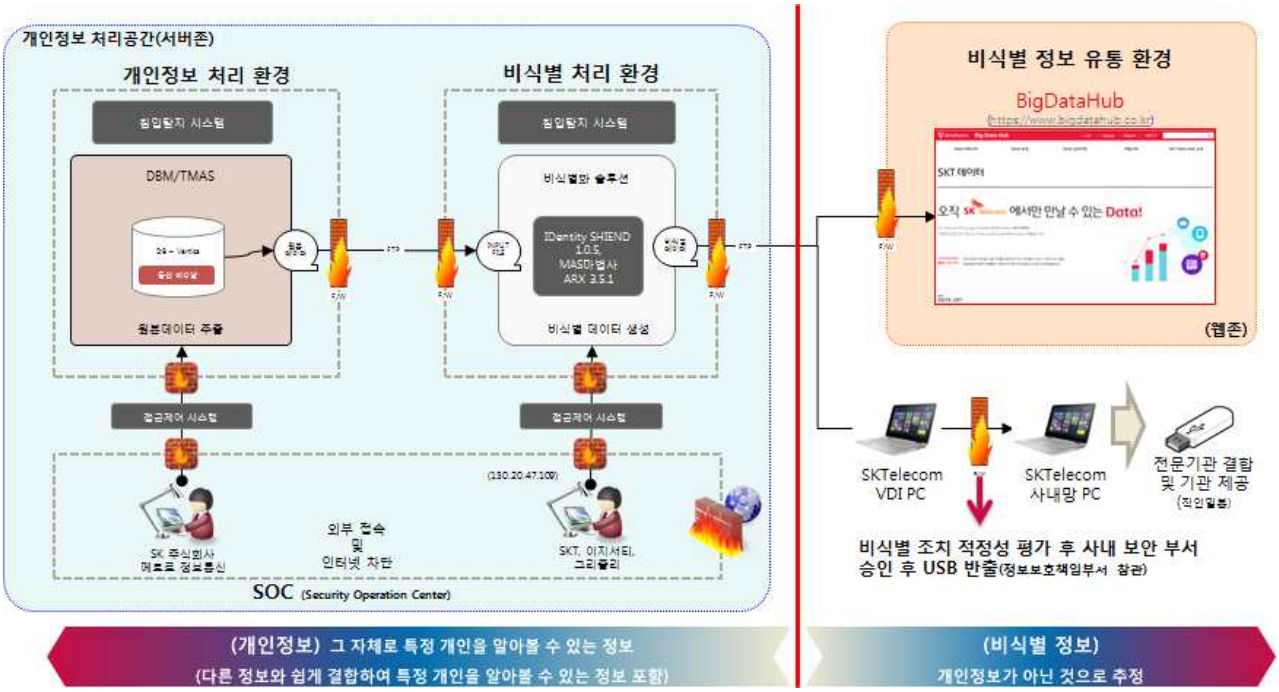
(5) 비식별용 데이터 추출 절차

비식별화 솔루션에 제공하는 데이터는 기존 DBM 데이터와 신규로 적재한 데이터 분석 통해 추출/가공하여 비식별화 서버에 제공을 함.

데이터 구성시 데이터간 상호 정합성을 고려하여 추출해야 정확한 데이터를 추출가능

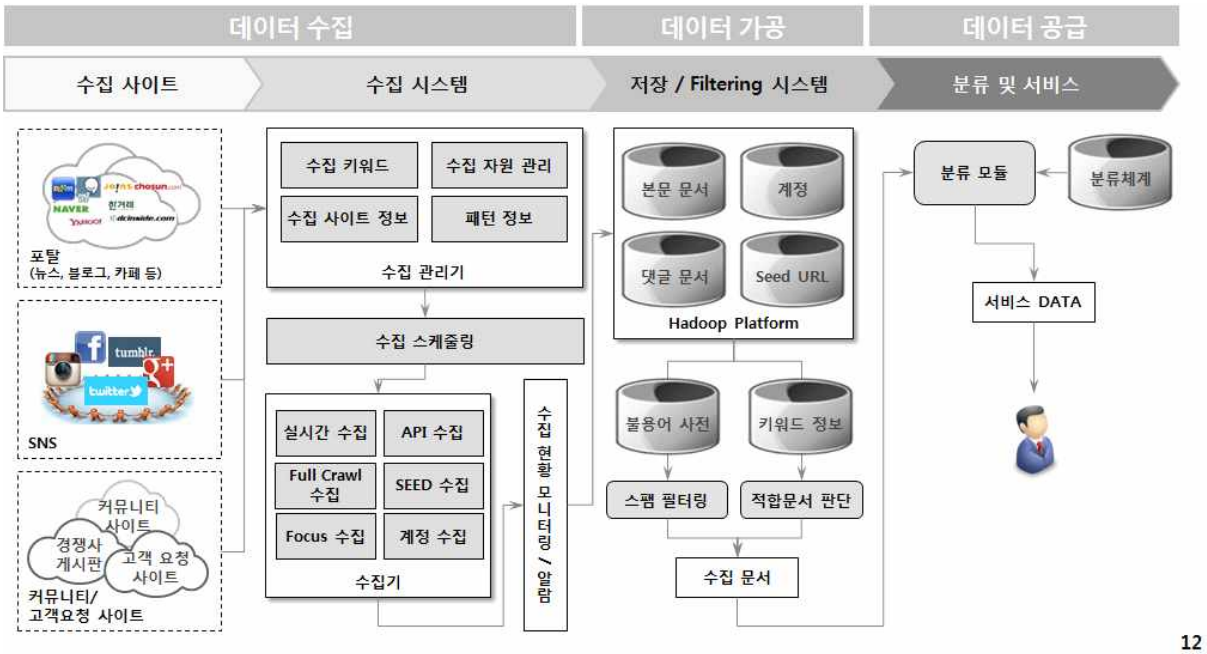


(6) 비식별화 시스템 구성도



(7) 외부 비정형 데이터 연계 시스템 구성

SNS 사이트의 open API 방식은 물론 HTML로 구성된 다양한 Web Site의 본문/댓글 등 다양한 수집원에서 발생하는 비정형 데이터 수집서비스를 제공.



2. 데이터 거래실증

가. 데이터 유통환경 구축

(1) Big Data Hub Data 카테고리 신규 구성

대분류 카테고리 5종, 중분류 카테고리 3종으로 Data 분류, 관리가능토록 하였음.

(2) Data 제공

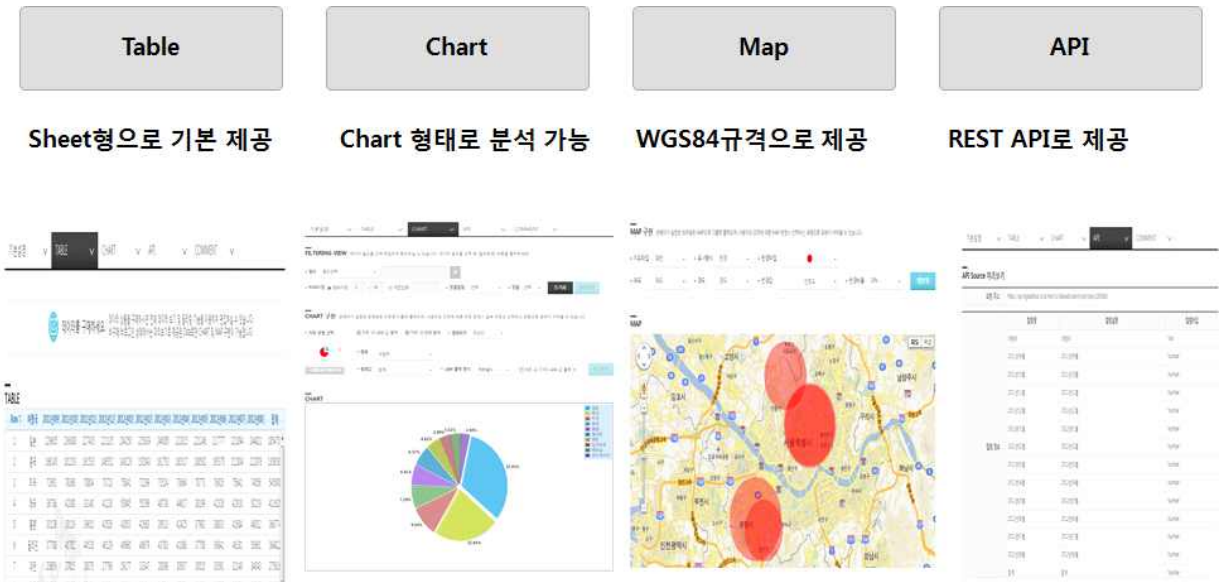
과제를 통해 생성된 SKT 프리미엄 Data, 화장품 10종 성분 Data, T-Raw Data(비식별화된 신용도 데이터, 사회약자 및 외국인 체류지 정보)를 제공환경 구성.

(3) Big Data Hub 커뮤니티

크롤링 뉴스와 사용자 문의에 대하여 전문가적인 답변을 관리하는 게시판 형태의 커뮤니티를 카테고리별로 제공하여 데이터 수요자와 공급자간의 의견활성화 지원

(4) Data 이용 및 체험환경 구성

Data에 따라 Table/Chart/Map/Download/API/Report 등 별도의 활용 제공 유형 없이 외부 Link 형태로 이용 및 분석

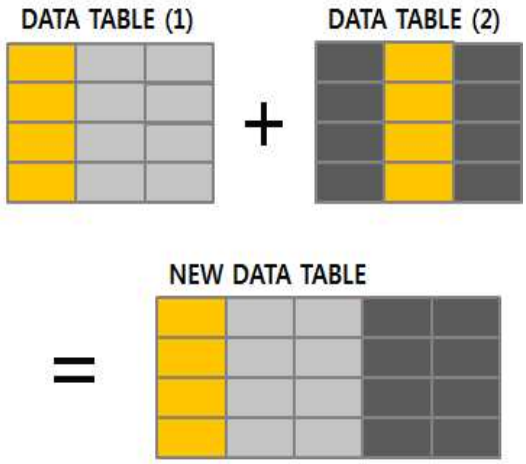


(5) Data 활용 API 제공

API 제공 상품 대상으로 Data 이용 신청을 통해 API 호출기반 정보 서비스 개발 및 활용이 가능

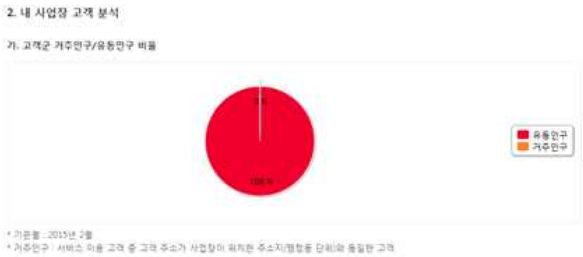
(6) Big Data Hub Data 매쉬업

사용자가 SKT만의 데이터와 내 데이터의 결합을 통하여 CHART, MAP 기능을 활용하여 다각도로 분석 환경 제공



(7) Big Data Hub Biz Call

통화량 데이터를 가공 및 분석한 자료로부터 생성된 보고서 제공



(8) 개발된 Big Data Hub 특장점

(가) 데이터를 공유, 거래하는 Open Data Marketplace

(나) SKT가 보유한 가입자 이용 Data를 기반으로 다양한 목적성 있는 데이터 제공

(다) 순수 국내 기술 비식별화 솔루션을 적용한 신용도 데이터, 사회약자 거소지 데이터, 외국인 체류지 정보 제공

(라) 사용자의 편의성을 위한 Data 등록관리, API사용 현황, 문의 내역 관리

(마) 스마트인사이트에서 제공하는 크롤링 뉴스, 화장품 10종 성분 Data 제공

(바) Data에 따라 TABLE, CHART, MAP, DOWN, API, Report 유형으로 제공

(사) 개발자에게 필요한 Data를 제공하고 API연동기능 제공

(아) Data 유형 중에서 TABLE 분석, CHART 분석, MAP 분석 방법 제공

(자) 이용자가 개발한 Application을 App 갤러리에 전시하는 공간 제공

(차) 통화량 데이터를 가공하여 소상공인 경영지원을 위한 Biz콜 자료 제공

나. H보험과 S신용평가 데이터 결합 실증

한화생명

- SKT 가입자 중 제3자 데이터 제공 동의를 한 1800만 대상, 한화 생명의 459만 데이터 연계 실증 연구 진행
- 2월 23일 KISA추천 적정정전문가Pool을 모시고 1차 적정성 평가 회의 진행 완료
- 3월 8일 KISA를 통한 데이터 연계 진행(한화생명 기준 40% 218만 데이터 연계 성공)
- 3월 3주차 연계 데이터 대상 2차 적절성 평가 회의

실증 결과

Alternative 신용평가 가능성 확인

결합 데이터만로 정확한 영향도를 확인할 순 없으나, 동일한 신용등급의 고객도 통신사와 생명보험의 연체정보만을 활용하여 신용대출 연체 발생 위험의 세분화 가능성을 일부 확인함

신용평가 확인대상 — 최근 5년 內 대출 後 현재까지 신용등급의 변화가 적은 고객 (123등급 → 123등급 유지, 456등급 → 456등급 유지)

123등급 유지고객 신용대출 연체율

구분	보험료납입 연체 경험	보험료납입 미연체
최근1년 통신료 연체	2.5%	1.1%
최근1년 통신료 미연체	1.4%	0.5%

최근1년 보험료납입 연체경험 時 5.9%

456등급 유지고객 신용대출 연체율

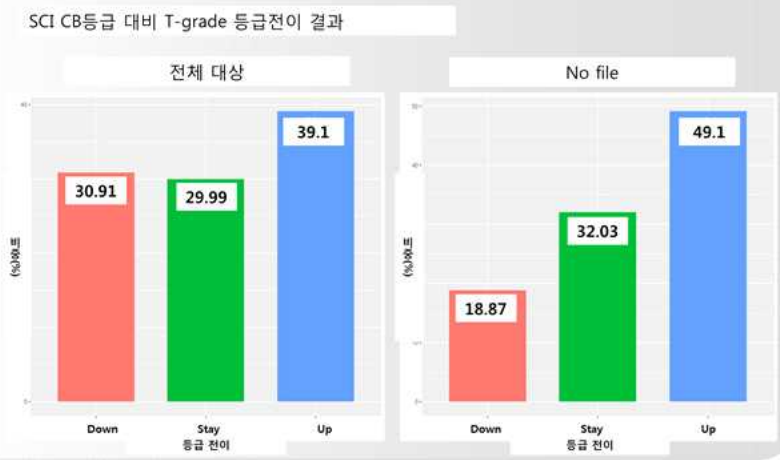
구분	보험료납입 연체 경험	보험료납입 미연체
최근1년 통신료 연체	6.5%	2.1%
최근1년 통신료 미연체	3.1%	1.0%

최근1년 보험료납입 연체경험 時 7.9%

SCI신용평가

- SKT 가입자 중 제3자 데이터 제공 동의를 한 서울 지역 가입자 대상 MAS 방법론을 통한 데이터 연계 실증 연구 진행
- 1차 데이터 비식별화 프로세싱
- 3월 3주차 데이터대상 적절성 평가 회의
- 3월 4주차 MAS 솔루션을 통한 데이터 연계 방법론 실증
- 실제 CB 스코어링의 적용을 통해 중금리 대출 이용자의 신용도 향상 가능성 검증

실증 결과



3. KLT 프라이버시 모델 실증(이지서티)

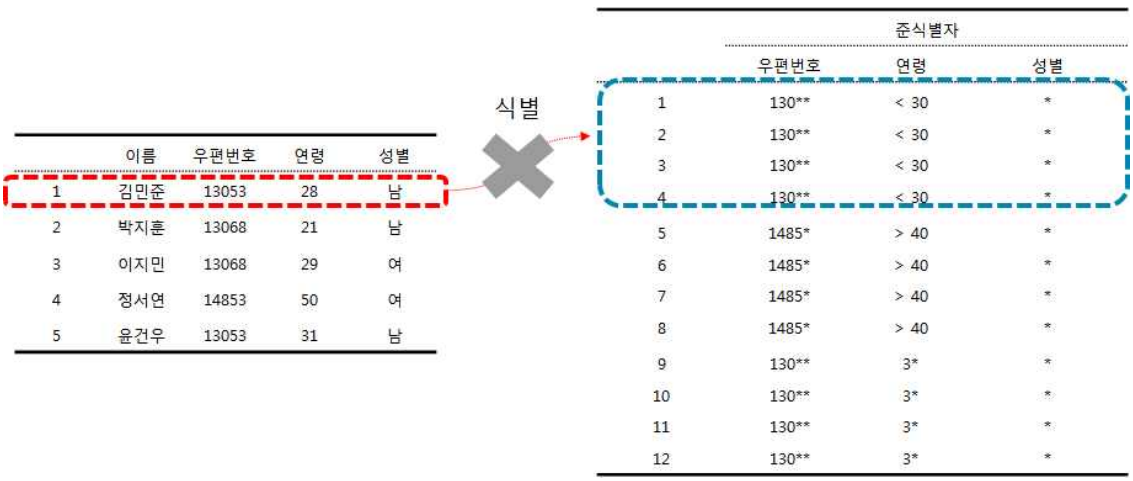
가. K. L. T 프라이버시 모델

(1) K. L. T 프라이버시 모델 개요

○ K. L. T 비식별 조치는 개인의 식별할 수 있는 식별자를 제거하여 개인을 식별할 수 없도록 조치하는 것을 비식별 조치라 함

○ K. L. T 는 k-익명성, l-다양성, t-근접성을 의미하며 K. L. T는 익명성, 다양성, 근접성에 대한 계수를 의미함

○ k-익명성: 데이터 집합에서 구별되지 않는 레코드(즉, 모든 속성이 동일한 레코드)를 최소 K개 이상으로 만들어 프라이버시를 보호하는 모델

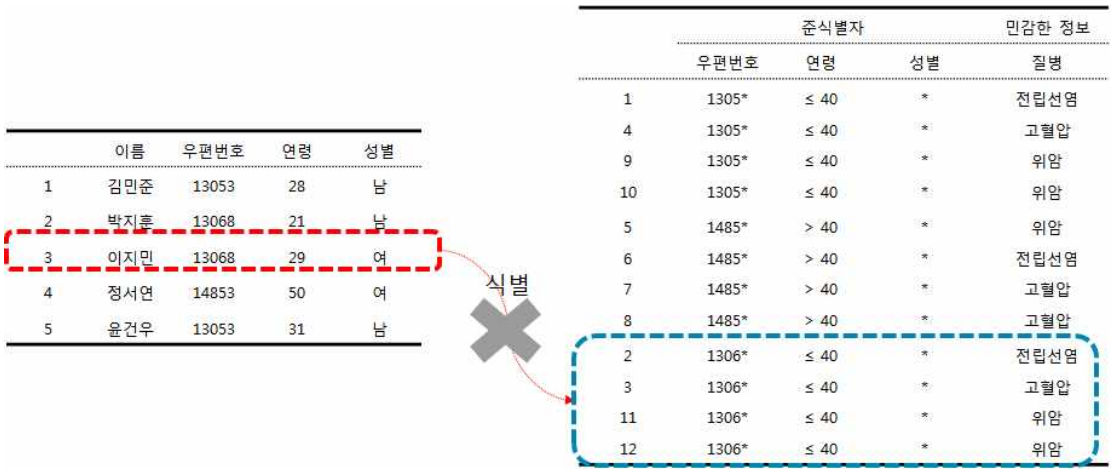


	이름	우편번호	연령	성별
1	김민준	13053	28	남
2	박지훈	13068	21	남
3	이지민	13068	29	여
4	정서연	14853	50	여
5	윤건우	13053	31	남

	준식별자		
	우편번호	연령	성별
1	130**	< 30	*
2	130**	< 30	*
3	130**	< 30	*
4	130**	< 30	*
5	1485*	> 40	*
6	1485*	> 40	*
7	1485*	> 40	*
8	1485*	> 40	*
9	130**	3*	*
10	130**	3*	*
11	130**	3*	*
12	130**	3*	*

[그림 3-1] k-익명성 예시

○ l-다양성: 데이터 집합에서 구별되지 않는 레코드들이 L개 이상의 민감한 정보(분석에 필요한 정보)를 가지도록 하여 프라이버시를 보호하는 모델



	이름	우편번호	연령	성별
1	김민준	13053	28	남
2	박지훈	13068	21	남
3	이지민	13068	29	여
4	정서연	14853	50	여
5	윤건우	13053	31	남

	준식별자			민감한 정보
	우편번호	연령	성별	질병
1	1305*	≤ 40	*	전립선염
4	1305*	≤ 40	*	고혈압
9	1305*	≤ 40	*	위암
10	1305*	≤ 40	*	위암
5	1485*	> 40	*	위암
6	1485*	> 40	*	전립선염
7	1485*	> 40	*	고혈압
8	1485*	> 40	*	고혈압
2	1306*	≤ 40	*	전립선염
3	1306*	≤ 40	*	고혈압
11	1306*	≤ 40	*	위암
12	1306*	≤ 40	*	위암

[그림 3-2] l-다양성 예시

○ t-근접성: 데이터 집합에서 구별되지 않는 레코드들의 민감한 정보의 분포와 전체 데이터

	준식별자		민감한 정보	
	우편번호	연령	급여 (백만원)	질병
1	4767*	≤ 40	30	위궤양
3	4767*	≤ 40	50	만성 위염
8	4767*	≤ 40	90	폐렴
4	4790*	≥ 40	60	급성 위염
5	4790*	≥ 40	110	감기
6	4790*	≥ 40	80	기관지염
2	4760*	3*	40	급성 위염
7	4760*	3*	70	기관지염
9	4760*	3*	100	만성 위염

[그림 3-3] t-근접성 예시

○ K. L. T를 이용하여 데이터 집합의 레코드를 동일한 레코드 K개 이상 K개의 레코드의 민감한 정보를 L개 만큼 다양하게 하고 민감한 데이터의 분포와 전체 데이터의 분포 차이를 T값(0~1사이) 이하로 만들어 프라이버시를 보호하는 비식별 조치 모델

나. SKT 3종 데이터 k-익명성 기법 비식별 조치

(1) SKT 3종 데이터 비식별 조치 개요

○ SKT의 20대 신용도 데이터, 장애우 거소지 데이터, 외국인 체류지 데이터를 비식별 조치 가이드라인의 비식별 조치 절차 및 요건을 준수 하여 비식별 대상 데이터의 전처리 수행, k-익명성의 수준 설정을 통한 비식별 조치를 수행하였음

<표 3-1> 비식별 대상 데이터 명세

구분	20대 신용도 데이터	장애우 거소지 데이터	외국인 체류지 데이터
컬럼 수	31	54	54
ROW 수	7,912,143	805,082	311,913
식별자 수	1	1	1
준 식별자 수	6	4	4
민감정보 수	24	49	49

(2) SKT 3종 데이터 비식별 조치 전처리

○ 20대 신용도, 장애우 거소지 데이터, 외국인 체류지 데이터에서 비식별 조치 목적에 따른 식별자, 준식별자 및 민감정보 등 데이터 속성 분류 및 전처리 방안 결정, 준식별자의 계층 트리 생성 등의 비식별 조치 전처리 수행

〈표 3-2〉 전처리 후 비식별 대상 데이터 속성 분류 정보

구분	20대 신용도 데이터	장애우 거소지 데이터	외국인 체류지 데이터
식별자 칼럼	0	0	0
준식별자 칼럼	6	4	3
민감정도 칼럼	18	48	48

20대 신용도-원본데이터		속성구분정보
순번	한글명	
1	기준 월	민감정보
2	가입자 이름	준식별자
3	이동전화번호	식별자
4	월 평균 통화시간	민감정보
5	월 평균 통화빈도	민감정보
6	서비스 개월 수	민감정보
7	가입자 생년월	준식별자
8	가입자 성별	준식별자
9	멤버쉽 등급	민감정보
10	멤버쉽 카드 발급 여부	민감정보
11	거주지역-시도	준식별자
12	거주지역-시군구	준식별자
13	거주지역-읍면동	준식별자
14	2nd Device 가입여부	민감정보
15	ARPU	민감정보
16	당월 납부요금	민감정보
17	결합상품 가입여부	민감정보
18	단말기 제조사	민감정보
19	단말기종	민감정보
20	단말기 출고가	민감정보
21	가입일자	민감정보
22	이용정지기간	민감정보
23	미납구분코드	민감정보
24	당월연체유무	민감정보
25	당월연체금액	민감정보
26	최근 1년간 납부일 미준수 횟수	민감정보
27	최근 1년간 최대 연체금액	민감정보
28	납부방법	민감정보
29	회선상태	민감정보
30	남은 할부 원금	민감정보
31	남은 할부 잔여 기간	민감정보

20대 신용도-비식별데이터	
순번	한글명
2	가입자 이름
4	월 평균 통화시간
5	월 평균 통화빈도
7	가입자 생년월
8	가입자 성별
9	멤버쉽 등급
11	거주지역-시도
12	거주지역-시군구
13	거주지역-읍면동
14	2nd Device 가입여부
15	ARPU
16	당월 납부요금
17	결합상품 가입여부
20	단말기 출고가
21	가입일자
22	이용정지기간
24	당월연체유무
25	당월연체금액
26	최근 1년간 납부일 미준수 횟수
27	최근 1년간 최대 연체금액
28	납부방법
29	회선상태
30	남은 할부 원금
31	남은 할부 잔여 기간

[그림 3-4] 20대 신용도 데이터 스키마 및 전처리 매핑 정보

장애우 거소지-원본데이터		속성구분정보
순번	한글명	
1	기준월	민감정보
2	가입자 이름	준식별자
3	이동전화번호	식별자
4	가입자 생년월	준식별자
5	가입자 성별	준식별자
6	장애인타입	준식별자
7	주중_06_09_방문지역-시도	민감정보
8	주중_06_09_방문지역-시군구	민감정보
9	주중_06_09_방문지역-읍면동	민감정보
10	시간	민감정보
11	주중_09_12_방문지역-시도	민감정보
12	주중_09_12_방문지역-시군구	민감정보
13	주중_09_12_방문지역-읍면동	민감정보
14	시간	민감정보
15	주중_12_14_방문지역-시도	민감정보
16	주중_12_14_방문지역-시군구	민감정보
17	주중_12_14_방문지역-읍면동	민감정보
18	시간	민감정보
19	주중_14_18_방문지역-시도	민감정보
20	주중_14_18_방문지역-시군구	민감정보
21	주중_14_18_방문지역-읍면동	민감정보
22	시간	민감정보
23	주중_18_22_방문지역-시도	민감정보
24	주중_18_22_방문지역-시군구	민감정보
25	주중_18_22_방문지역-읍면동	민감정보
26	시간	민감정보
27	주중_22_06_방문지역-시도	민감정보
28	주중_22_06_방문지역-시군구	민감정보
29	주중_22_06_방문지역-읍면동	민감정보
30	시간	민감정보
31	주말_06_09_방문지역-시도	민감정보
32	주말_06_09_방문지역-시군구	민감정보
33	주말_06_09_방문지역-읍면동	민감정보
34	시간	민감정보
35	주말_09_12_방문지역-시도	민감정보
36	주말_09_12_방문지역-시군구	민감정보
37	주말_09_12_방문지역-읍면동	민감정보
38	시간	민감정보
39	주말_12_14_방문지역-시도	민감정보
40	주말_12_14_방문지역-시군구	민감정보
41	주말_12_14_방문지역-읍면동	민감정보
42	시간	민감정보
43	주말_14_18_방문지역-시도	민감정보
44	주말_14_18_방문지역-시군구	민감정보
45	주말_14_18_방문지역-읍면동	민감정보
46	시간	민감정보
47	주말_18_22_방문지역-시도	민감정보
48	주말_18_22_방문지역-시군구	민감정보
49	주말_18_22_방문지역-읍면동	민감정보
50	시간	민감정보
51	주말_22_06_방문지역-시도	민감정보
52	주말_22_06_방문지역-시군구	민감정보
53	주말_22_06_방문지역-읍면동	민감정보
54	시간	민감정보

장애우 거소지-비식별데이터	
순번	한글명
2	가입자 이름
4	가입자 생년월
5	가입자 성별
6	장애인타입
7	주중_06_09_방문지역-시도
8	주중_06_09_방문지역-시군구
9	주중_06_09_방문지역-읍면동
10	시간
11	주중_09_12_방문지역-시도
12	주중_09_12_방문지역-시군구
13	주중_09_12_방문지역-읍면동
14	시간
15	주중_12_14_방문지역-시도
16	주중_12_14_방문지역-시군구
17	주중_12_14_방문지역-읍면동
18	시간
19	주중_14_18_방문지역-시도
20	주중_14_18_방문지역-시군구
21	주중_14_18_방문지역-읍면동
22	시간
23	주중_18_22_방문지역-시도
24	주중_18_22_방문지역-시군구
25	주중_18_22_방문지역-읍면동
26	시간
27	주중_22_06_방문지역-시도
28	주중_22_06_방문지역-시군구
29	주중_22_06_방문지역-읍면동
30	시간
31	주말_06_09_방문지역-시도
32	주말_06_09_방문지역-시군구
33	주말_06_09_방문지역-읍면동
34	시간
35	주말_09_12_방문지역-시도
36	주말_09_12_방문지역-시군구
37	주말_09_12_방문지역-읍면동
38	시간
39	주말_12_14_방문지역-시도
40	주말_12_14_방문지역-시군구
41	주말_12_14_방문지역-읍면동
42	시간
43	주말_14_18_방문지역-시도
44	주말_14_18_방문지역-시군구
45	주말_14_18_방문지역-읍면동
46	시간
47	주말_18_22_방문지역-시도
48	주말_18_22_방문지역-시군구
49	주말_18_22_방문지역-읍면동
50	시간
51	주말_22_06_방문지역-시도
52	주말_22_06_방문지역-시군구
53	주말_22_06_방문지역-읍면동
54	시간

[그림 3-5] 장애우 거소지 데이터 스키마 및 전처리 매핑 정보

외국인 체류지-원본데이터		속성구분정보
순번	한글명	
1	기준월	민감정보
2	가입자 이름	준식별자
3	이동전화번호	식별자
4	가입자 생년월	준식별자
5	가입자 성별	준식별자
6	외국인 장애인타입	준식별자
7	주중_06_09_방문지역-시도	민감정보
8	주중_06_09_방문지역-시군구	민감정보
9	주중_06_09_방문지역-읍면동	민감정보
10	시간	민감정보
11	주중_09_12_방문지역-시도	민감정보
12	주중_09_12_방문지역-시군구	민감정보
13	주중_09_12_방문지역-읍면동	민감정보
14	시간	민감정보
15	주중_12_14_방문지역-시도	민감정보
16	주중_12_14_방문지역-시군구	민감정보
17	주중_12_14_방문지역-읍면동	민감정보
18	시간	민감정보
19	주중_14_18_방문지역-시도	민감정보
20	주중_14_18_방문지역-시군구	민감정보
21	주중_14_18_방문지역-읍면동	민감정보
22	시간	민감정보
23	주중_18_22_방문지역-시도	민감정보
24	주중_18_22_방문지역-시군구	민감정보
25	주중_18_22_방문지역-읍면동	민감정보
26	시간	민감정보
27	주중_22_06_방문지역-시도	민감정보
28	주중_22_06_방문지역-시군구	민감정보
29	주중_22_06_방문지역-읍면동	민감정보
30	시간	민감정보
31	주말_06_09_방문지역-시도	민감정보
32	주말_06_09_방문지역-시군구	민감정보
33	주말_06_09_방문지역-읍면동	민감정보
34	시간	민감정보
35	주말_09_12_방문지역-시도	민감정보
36	주말_09_12_방문지역-시군구	민감정보
37	주말_09_12_방문지역-읍면동	민감정보
38	시간	민감정보
39	주말_12_14_방문지역-시도	민감정보
40	주말_12_14_방문지역-시군구	민감정보
41	주말_12_14_방문지역-읍면동	민감정보
42	시간	민감정보
43	주말_14_18_방문지역-시도	민감정보
44	주말_14_18_방문지역-시군구	민감정보
45	주말_14_18_방문지역-읍면동	민감정보
46	시간	민감정보
47	주말_18_22_방문지역-시도	민감정보
48	주말_18_22_방문지역-시군구	민감정보
49	주말_18_22_방문지역-읍면동	민감정보
50	시간	민감정보
51	주말_22_06_방문지역-시도	민감정보
52	주말_22_06_방문지역-시군구	민감정보
53	주말_22_06_방문지역-읍면동	민감정보
54	시간	민감정보

외국인 체류지-비식별데이터	
순번	한글명
2	가입자 이름
4	가입자 생년월
5	가입자 성별
7	주중_06_09_방문지역-시도
8	주중_06_09_방문지역-시군구
9	주중_06_09_방문지역-읍면동
10	시간
11	주중_09_12_방문지역-시도
12	주중_09_12_방문지역-시군구
13	주중_09_12_방문지역-읍면동
14	시간
15	주중_12_14_방문지역-시도
16	주중_12_14_방문지역-시군구
17	주중_12_14_방문지역-읍면동
18	시간
19	주중_14_18_방문지역-시도
20	주중_14_18_방문지역-시군구
21	주중_14_18_방문지역-읍면동
22	시간
23	주중_18_22_방문지역-시도
24	주중_18_22_방문지역-시군구
25	주중_18_22_방문지역-읍면동
26	시간
27	주중_22_06_방문지역-시도
28	주중_22_06_방문지역-시군구
29	주중_22_06_방문지역-읍면동
30	시간
31	주말_06_09_방문지역-시도
32	주말_06_09_방문지역-시군구
33	주말_06_09_방문지역-읍면동
34	시간
35	주말_09_12_방문지역-시도
36	주말_09_12_방문지역-시군구
37	주말_09_12_방문지역-읍면동
38	시간
39	주말_12_14_방문지역-시도
40	주말_12_14_방문지역-시군구
41	주말_12_14_방문지역-읍면동
42	시간
43	주말_14_18_방문지역-시도
44	주말_14_18_방문지역-시군구
45	주말_14_18_방문지역-읍면동
46	시간
47	주말_18_22_방문지역-시도
48	주말_18_22_방문지역-시군구
49	주말_18_22_방문지역-읍면동
50	시간
51	주말_22_06_방문지역-시도
52	주말_22_06_방문지역-시군구
53	주말_22_06_방문지역-읍면동
54	시간

[그림 3-6] 외국인 체류지 데이터 스키마 및 전처리 매핑 정보

(3) SKT 3종 데이터 비식별 조치 수행 및 결과

○ SKT 3종 데이터의 비식별 조치 목적에 따라 선정된 준식별자, 민감정보 속성의 k-익명성, l-다양성 값의 수준을 설정하여 비식별 조치 수행

○ SKT 3종 데이터의 비식별 조치 수준은 k-익명성 수준 3, 5로 설정하여 각 데이터 마다 총 2회 진행 l-다양성은 미설정

<표 3-3> SKT 3종 데이터 비식별 조치 결과

구분	20대 신용도 데이터		장애우 거소지 데이터		외국인 체류지 데이터	
비식별 수준 설정 값	K=3	K=5	K=3	K=5	K=3	K=5
비식별 수준 결과 값	K=8222	K=5	K=14	K=36	K=22	K=120

※ 20대 신용도 약 800만건 데이터의 준식별자 컬럼의 구간화 전처리로 인한 k-익명성 결과 수준의 증가

다. H보험 외부기관 데이터 결합

(1) H보험 외부기관 데이터 결합 개요

○ SKT의 데이터 총 18,019,816건과 H보험의 데이터 결합을 통해 20대 고객층의 신용도를 책정할 수 있는 지에 대한 분석을 목적으로 비식별조치 가이드라인의 절차 및 가이드에 따라 비식별 조치 및 외부기관 데이터 결합을 진행

(2) SKT 고객 데이터 k-익명성, l-다양성 비식별 조치

○ 비식별 대상 데이터 스키마

* 총 23 칼럼 중 식별자 2개, 준식별자 2개, 민감정도 19개 컬럼으로 분류

SK텔레콤 고객 데이터-원본데이터			속성구분정보
순번	한글명	데이터 예시	
1	이름	홍길동	식별자
2	주민번호앞7자리	7107181	식별자
3	나이	35	준식별자
4	성별	1	준식별자
5	사용개월수	3년이상	민감정보
6	멤버쉽등급	V	민감정보
7	월평균통화시간	74	민감정보
8	월평균통화빈도	101	민감정보
9	ARPU	61730	민감정보
10	결합상품가입여부	Y	민감정보
11	단말기출고가	599999	민감정보
12	이용정지기간_개월	1개월미만	민감정보
13	당월연체금액	3000000	민감정보
14	최근1년간최대연체금액	3500000	민감정보
15	납부방법	카드자동납부	민감정보
16	회선상태	사용중	민감정보
17	남은할부원금	100000	민감정보
18	가입회선수	2	민감정보
19	Tablet 보유여부	Y	민감정보
20	Smart Watch 보유여부	N	민감정보
21	멤버쉽 월 사용금액	1799	민감정보
22	멤버쉽 년 사용금액	53079	민감정보
23	미납횟수	1	민감정보

[그림 3-7] SKT 데이터 스키마 구분

(3) SKT 고객 데이터 k-익명성, l-다양성 비식별 조치 수행 및 결과

○ 식별자 : 이름, 주민번호앞 7자리 (2개 컬럼) → 삭제

○ 준식별자 : 나이, 성별 (2개 컬럼) → 범주화

○ 민감정보 : 사용개월수, 맴버쉽등급, 월평균통화시간, 월평균통화빈도 등 (19개 칼럼) → 범주화(13개 컬럼), 데이터마스킹(1개 컬럼), 미적용(5개 컬럼)

SK텔레콤 고객 데이터-원본데이터			속성구분정보
순번	한글명	적용된 비식별 조치 기법	
1	이름	암호화(임시대체키 생성	식별자
2	주민번호앞7자리		식별자
3	나이	범주화(범위방법)	준식별자
4	성별	범주화(범위방법)	준식별자
5	사용개월수	범주화(범위방법)	민감정보
6	멤버쉽등급	데이터마스킹(대체)	민감정보
7	월평균통화시간	범주화(범위방법)	민감정보
8	월평균통화빈도	범주화(범위방법)	민감정보
9	ARPU	범주화(범위방법)	민감정보
10	결합상품가입여부	미적용	민감정보
11	단말기출고가	범주화(범위방법)	민감정보
12	이용정지기간_개월	범주화(범위방법)	민감정보
13	당월연체금액	범주화(범위방법)	민감정보
14	최근1년간최대연체금액	범주화(범위방법)	민감정보
15	납부방법	미적용	민감정보
16	회선상태	미적용	민감정보
17	남은할부원금	범주화(범위방법)	민감정보
18	가입회선수	범주화(범위방법)	민감정보
19	Tablet 보유여부	미적용	민감정보
20	Smart Watch 보유여부	미적용	민감정보
21	멤버쉽 월 사용금액	범주화(범위방법)	민감정보
22	멤버쉽 년 사용금액	범주화(범위방법)	민감정보
23	미납횟수	범주화(범위방법)	민감정보

[그림 3-8] 비식별 조치 알고리즘 설정

〈표 3-4〉 SKT 고객 데이터 비식별 조치 결과

구분	H보험 데이터 결합용 SKT 고객 데이터			
비식별 수준 설정 값	k-익명성 수준	K=3	l-다양성 수준	L=2 (당월 연체 금액)
비식별 수준 결과 값	k-익명성 수준	K=114463	l-다양성 수준	L=8

※ 약 1,800만건 데이터의 준식별자 컬럼인 나이(18~61세) 및 성별(남, 여)의 구간화 전처리로 인한 k-익명성 결과 수준의 증가

(4) SKT 과 H보험 결합 데이터 비식별 조치

○ 비식별 대상 데이터 스키마 및 비식별 조치 적용 내용

SK&한화 -원본데이터			속성구분정보
순번	한글명	적용된 비식별 조치 기법	
1	나이	범주화(범위방법)	준식별자
2	성별	범주화(범위방법)	준식별자
3	직업	범주화(범위방법)	준식별자
4	신용대출 건수	범주화(범위방법)	속성(민감) 정보
5	최초 계약월	미적용	속성(민감) 정보
6	최초 연체 등록월	미적용	속성(민감) 정보
7	총 신용대출 금액_백만	범주화(범위방법)	속성(민감) 정보
8	총 상환 금액_백만	범주화(범위방법)	속성(민감) 정보
9	신용대출 연체율	미적용	속성(민감) 정보
10	30일이상 신용대출 연체율	미적용	속성(민감) 정보
11	최근1년 신용대출 연체율	미적용	속성(민감) 정보
12	최초 신용 등급	미적용	속성(민감) 정보
13	최근 신용 등급	미적용	속성(민감) 정보
14	보험료 연체율	미적용	속성(민감) 정보
15	최근 1년 보험료 연체율	미적용	속성(민감) 정보
16	실효 해지 건수	범주화(범위방법)	속성(민감) 정보
17	기납입 보험료	범주화(범위방법)	속성(민감) 정보
18	납입 보험료	범주화(범위방법)	속성(민감) 정보
19	추정소득_백만	범주화(범위방법)	속성(민감) 정보
20	가구 추정 소득	범주화(범위방법)	속성(민감) 정보
21	평균약관 대출율	미적용	속성(민감) 정보
22	약관 대출 금액_십만	범주화(범위방법)	속성(민감) 정보
23	자동이체 실패월수	범주화(범위방법)	속성(민감) 정보
24	사용개월수	범주화(범위방법)	속성(민감) 정보
25	멤버쉽등급	범주화(범위방법)	속성(민감) 정보
26	월평균통화시간	범주화(범위방법)	속성(민감) 정보
27	월평균통화빈도	범주화(범위방법)	속성(민감) 정보
28	가입자당 평균 매출	범주화(범위방법)	속성(민감) 정보
29	결합상품가입여부	미적용	속성(민감) 정보
30	단말기출고가	범주화(범위방법)	속성(민감) 정보
31	이용정지기간_개월	범주화(범위방법)	속성(민감) 정보
32	당월연체금액	범주화(범위방법)	속성(민감) 정보
33	최근1년간최대연체금액	범주화(범위방법)	속성(민감) 정보
34	납부방법	미적용	속성(민감) 정보
35	회선상태	미적용	속성(민감) 정보
36	남은할부원금	범주화(범위방법)	속성(민감) 정보
37	가입 회선수	범주화(범위방법)	속성(민감) 정보
38	Tablet 보유여부	미적용	속성(민감) 정보
39	Smart Watch 보유여부	미적용	속성(민감) 정보
40	멤버쉽 월 사용금액	범주화(범위방법)	속성(민감) 정보
41	멤버쉽 년 사용금액	범주화(범위방법)	속성(민감) 정보
42	미납횟수	범주화(범위방법)	속성(민감) 정보

[그림 3-9] SKT과 H보험 결합 데이터 스키마 및 비식별 조치 알고리즘 설정

(5) SKT 고객 데이터 k-익명성, l-다양성 비식별 조치 수행 및 결과

○ 준식별자 : 나이, 성별, 직업 (2개 컬럼) → 범주화

○ 민감정보 : 신용대출 건수, 총 신용대출 금액, 총 상환 금액, 실효 해지 건수, 기납입 보험료 등 (39개 컬럼) → 범주화(24개 컬럼), 미적용(15개 컬럼)

〈표 3-5〉 SKT 및 H보험 결합 데이터 비식별 조치 결과

구분	H보험 데이터 결합용 SKT 고객 데이터			
비식별 수준 설정 값	k-익명성 수준	K=5	l-다양성 수준	L=2 (당월 연체 금액)
비식별 수준 결과 값	k-익명성 수준	K=5	l-다양성 수준	L=2

라. 데이터 결합 검증용 데이터 MASH-UP

(1) SKT 데이터 MASH-UP 개요

○ SKT 20대 신용도 데이터의 29개 칼럼을 MASH-UP 데이터 A, MASH-UP 데이터 B로 이분화 하여 두 데이터를 비식별 조치 이후 결합을 진행하여 결합도 검증 진행

(2) SKT MASH-UP 데이터 k-익명성, l-다양성 비식별 조치

○ 비식별 대상 데이터 스키마 및 비식별 조치 적용 내용

20대 신용도-MESHUP원본데이터			속성구분정보
순번	한글명	적용된 비식별 조치 기법	
1	열번호	미적용	결합용
2	이동전화번호	삭제	식별자
3	가입자 명	범주화	준식별자
4	가입자 생년월	범주화	준식별자
5	가입자 성별	범주화	준식별자
6	거주지역-시도	범주화	준식별자
7	거주지역-시군구		준식별자
8	거주지역-읍면동		준식별자
9	월 평균 통화시간	미적용	민감정보
10	월 평균 통화빈도	미적용	민감정보
11	월 평균 통화빈도	미적용	민감정보
12	ARPU	미적용	민감정보
13	당월 납부요금	미적용	민감정보
14	결합상품 가입여부	미적용	민감정보
15	단말기 출고가	미적용	민감정보
16	가입일자	미적용	민감정보
17	납부방법	미적용	민감정보
18	남은 할부 원금	미적용	민감정보
19	남은 할부 잔여 기간	미적용	민감정보

[그림 3-10] MASH-UP 데이터 A

20대 신용도-MESHUP원본데이터			속성구분정보
순번	한글명	적용된 비식별 조치 기법	
1	열번호	미적용	결합용
2	이동전화번호	삭제	식별자
3	가입자 명	범주화	준식별자
4	가입자 생년월	범주화	준식별자
5	가입자 성별	범주화	준식별자
6	거주지역-시도	범주화	준식별자
7	거주지역-시군구		준식별자
8	거주지역-읍면동		준식별자
9	멤버쉽사용금액	미적용	민감정보
10	당년멤버쉽사용금액	미적용	민감정보
11	멤버쉽 등급	미적용	민감정보
12	2nd Device 가입여부	미적용	민감정보
13	이용정지기간	미적용	민감정보
14	당월연체유무	미적용	민감정보
15	당월연체금액	미적용	민감정보
16	최근 1년간 납부일 미준수 횟수	미적용	민감정보
17	최근 1년간 최대 연체금액	미적용	민감정보
18	회선상태	미적용	민감정보

[그림 3-11] MASH-UP 데이터 B

(3) SKT MASH-UP 데이터 k-익명성, l-다양성 비식별 조치 수행 및 결과

○ 식별자 : 이동전화번호(1개 칼럼) → 삭제

○ 준식별자 : 가입자 명, 가입자 생년월, 가입자 성별, 거주지역 (4개 컬럼) → 범주화

○ 민감정보 : 월 평균 통화 시간, 월 평균 통화 빈도, 맴버쉽 사용금액, 당년 맵버쉽 사용금액 등 (21개 컬럼) → 미적용(21개 컬럼)

〈표 3-6〉 SKT MASH-UP 데이터 비식별 조치 결과

구분		SKT MASH-UP 데이터			
MASH_UP 데이터 A	비식별 수준 설정 값	k-익명성 수준	K=3	l-다양성 수준	L=2
	비식별 수준 결과 값	k-익명성 수준	K=3	l-다양성 수준	L=2
MASH_UP 데이터 B	비식별 수준 설정 값	k-익명성 수준	K=3	l-다양성 수준	L=2
	비식별 수준 결과 값	k-익명성 수준	K=3	l-다양성 수준	L=2

(4) SKT MASH-UP 데이터 결합

○ SKT MASH-UP 데이터 A, B의 대체 식별자를 기준으로 두 데이터를 결합진행 결합은 LEFT OUTER JOIN을 이용하여 결합

20대 신용도-MESHUP원본데이터			속성구분정보
순번	구분	한글명	
1	MESH-UP 데이터 A	가입자 명	준식별자
2		가입자 생년월	준식별자
3		가입자 성별	준식별자
4		거주지역-시도	준식별자
5		거주지역-시군구	준식별자
6		거주지역-읍면동	준식별자
7		멤버쉽사용금액	민감정보
8		당년멤버쉽사용금액	민감정보
9		멤버쉽 등급	민감정보
10		2nd Device 가입여부	민감정보
11		이용정지기간	민감정보
12		당월연체유무	민감정보
13		당월연체금액	민감정보
14		최근 1년간 납부일 미준수 횟수	민감정보
15		최근 1년간 최대 연체금액	민감정보
16		회선상태	민감정보
17	MESH-UP 데이터 B	가입자 명	준식별자
18		가입자 생년월	준식별자
19		가입자 성별	준식별자
20		거주지역-시도	준식별자
21		거주지역-시군구	준식별자
22		거주지역-읍면동	준식별자
23		월 평균 통화시간	민감정보
24		월 평균 통화빈도	민감정보
25		월 평균 통화빈도	민감정보
26		ARPU	민감정보
27		당월 납부요금	민감정보
28		결합상품 가입여부	민감정보
29		단말기 출고가	민감정보
30		가입일자	민감정보
31		납부방법	민감정보
32		남은 할부 원금	민감정보
33		남은 할부 잔여 기간	민감정보

[그림 3-12] 결합된 MASH-UP 데이터 A, B

마. k-익명성 기법 고도화

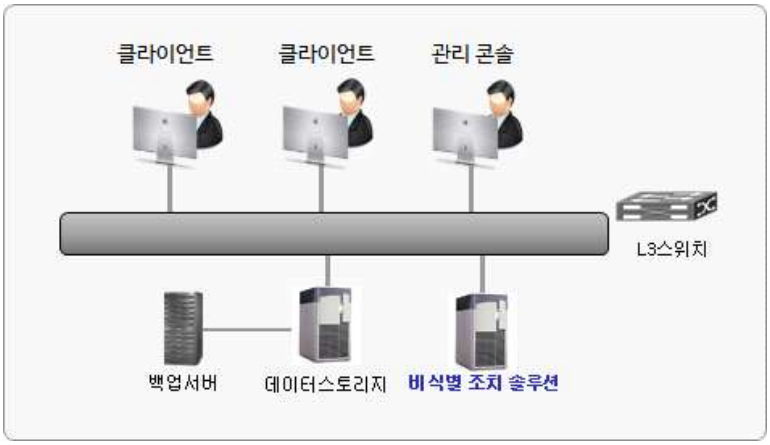
(1) k-익명성 기법 개요

○ 프라이버시 모델 비식별화에서 k-익명성은 준식별자(qi)에 해당하는 속성의 값이 동일한 Recode가 적어도 K개 이상이 있어야 k-익명성을 만족하며, k-익명성을 구하는 비식별화 알고리즘은 계층트리를 기준으로 계층 격자를 생성하고 계층 격자의 0단계부터 최종 단계까지 단계별로 k-익명성 값의 충족여부를 검사하는 과정을 반복 진행하여 대상 데이터의 k-익명성을 만족하는 비식별화를 진행

(2) k-익명성 기법 고도화

○ 알고리즘 고도화 테스트 환경

* 대용량 빅데이터를 이용하여 고도화된 기법 자체 테스트 진행

구분	내용
CPU	2 X Intel® Xeon® E7-4809v3 (2.0GHz/8-core/20MB/115W) Processor (16core)
MEMORY	128GB (8x16GB) PC4-2133P-R DIMMs (DDR4)
HDD	HP 600 GB 12G SAS 15K
테스트 데이터 명세	2,000만 건, 33개 칼럼으로 이루어진 임의 생성 테스트 데이터
시험 환경	

○ 계층트리를 기준으로 계층격자를 생성하고 계층 격자의 0단계부터 최종단계까지 단계별로 탐색하여 k-익명성 값의 충족 여부를 검사하는 과정을 진행하는 중간에 k-익명성 수준 이상을 만족하는 Recode 집합을 데이터 셋에서 추출하여 결과데이터로 분리하고 k-익명성 수준을 만족하지 못하는 Recode 집합을 계층격자의 다음단계를 탐색하는 위의 과정을 반복

(3) 기존 방식의 알고리즘 프로세스

① 준식별자 범주 트리로부터 계층 트리 및 계층격자 생성

② 격자의 가장 하위 노드부터 너비 우선탐색

③ 각 노드 일반화 조건에 따라 데이터 변환

④-Ⓐ k-익명성 만족시 비식별화 완료

④-Ⓑ k-익명성 불만족시 ②번 절차부터 다시 시작

(4) 고도화된 방식의 알고리즘 프로세스

① 준식별자 범주 트리로부터 계층 트리 및 계층격자 생성

② 격자의 가장 하위 노드부터 너비 우선탐색

③ 각 노드 일반화 조건에 따라 데이터 변환

④-Ⓐ k-익명성 만족시 비식별화 완료

④-Ⓑ k-익명성 불만족시 k-익명성을 만족하는 Recode 추출

⑤ 비식별화 미완료된 레코드가 없을 때까지 ②~④ 반복

(5) k-익명성 기법 알고리즘 고도화 결과

○ 알고리즘 고도화에 따라 전체 레코드를 계층격자를 탐색하며, k-익명성을 검사하던 방식에서 계층격자를 탐색하여 k-익명성을 만족하는 레코드를 추출하고 또 하위 계층격자에서 많은 데이터를 추출하기 때문에 데이터의 해상도 및 성능 면에서 향상됨

바. 외부기관 인증

(1) CC 인증

구분	내용
신청기관	주식회사 이지서티
인증기관	한국기계전기전자시험연구원
신청날짜	2017년 3월 27일
제품분류 / 신청등급	네트워크 개인정보 / EAL 2
시험검사항목	CC v3.1 R2 및 네트워크 개인정보보호제품 시험기준 V2.0/정보보호제품 시험.평가
시험 항목	1. 개인정보 검출 및 정보흐름통제
	2. 전송 데이터 보호
	3. 식별 및 인증
	4. 안전한 세션 관리
	5. 개인정보보호 서버와 업데이트 서버 간 안전한 연동
	6. 감사 기록
	7. 보안관리
	8. 자체시험

※ 현, 비식별조치에 대한 시험기준이 없어 개인정보보호 제품 시험기준을 따라 시험 진행중

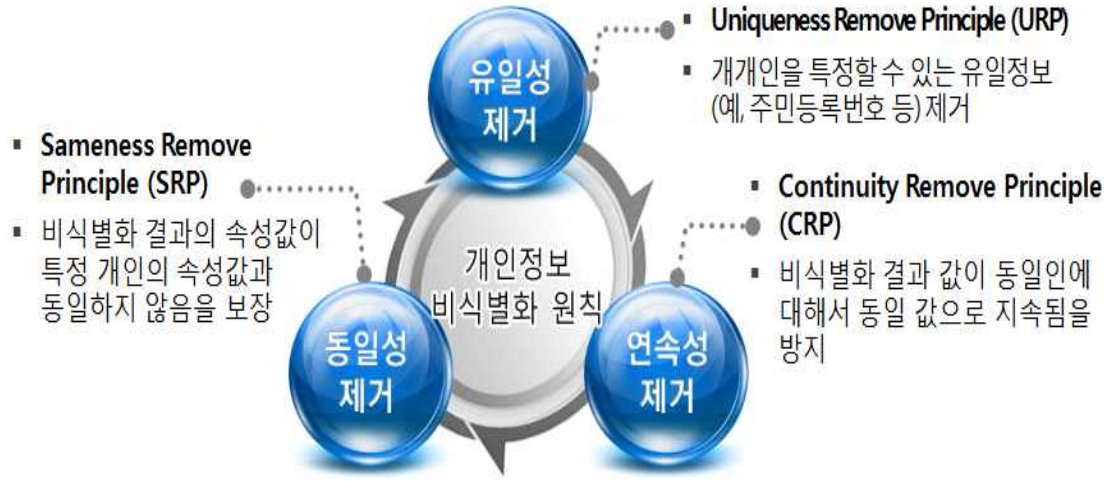
(2) V&V 인증

구분	내용	
신청기관	주식회사 이지서티	
인증기관	한국정보통신기술협회(TTA)	
신청날짜	2017년 4월 13일	
시험검사항목	아이덴티티 쉴드 v2.0 Verification & Validation	
시험 항목	1. 고전적 비식별조치 기능	데이터 업로드→비식별 조치 방법 선택→업로드 데이터의 개인정보 유형선택 및 고전적 비식별 알고리즘 선택→비식별화 실행→각 컬럼별 비식별 조치 결과 확인
	2. k-익명성 비식별조치 기능	데이터 업로드→비식별 조치 방법 선택→업로드 데이터의 개인정보 유형선택 및 비식별 알고리즘 선택→K 익명성 값 설정→비식별화 실행→ k 값 확인(입력 K 값 〈 결과 K값)
	3. l-다양성 비식별조치 기능	데이터 업로드→비식별 조치 방법 선택업로드 데이터의 개인정보 유형선택 및 비식별 알고리즘 선택→ k값 설정 및 L 다양성 값 설정→비식별화 실행→L 값 확인(입력 L 값 〈 결과 L값)

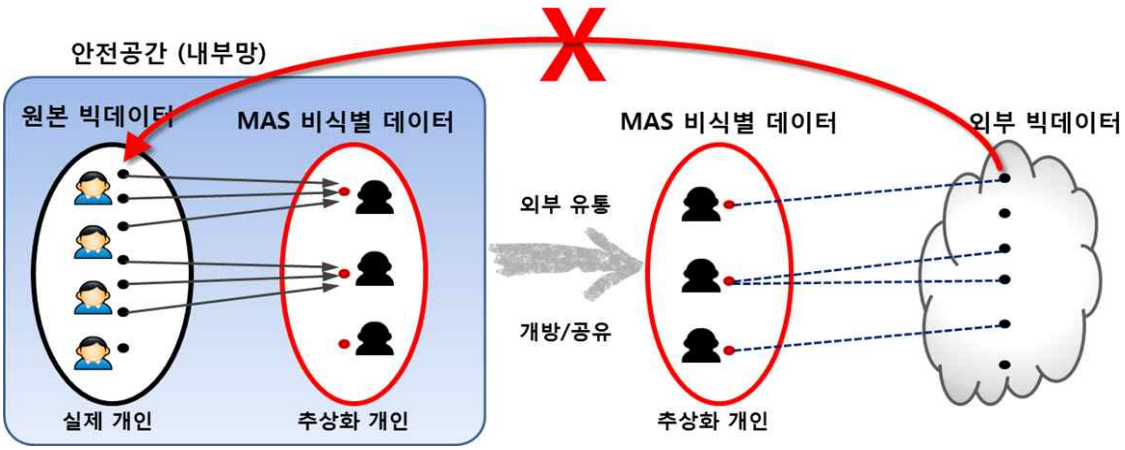
4. MAS 비식별화 알고리즘 실증(그리즐리)

가. 솔루션 개요

(1) 다수준 추상 동기화



○ 원본 빅데이터의 실제 개인들의 정보들에 대해 개인정보 비식별화 3대 원칙에 기반하여 유일성, 동일성, 연속성을 제거하고 실제 개인들의 정보를 변환 추상화하여 개인 재식별이 원천적으로 불가능



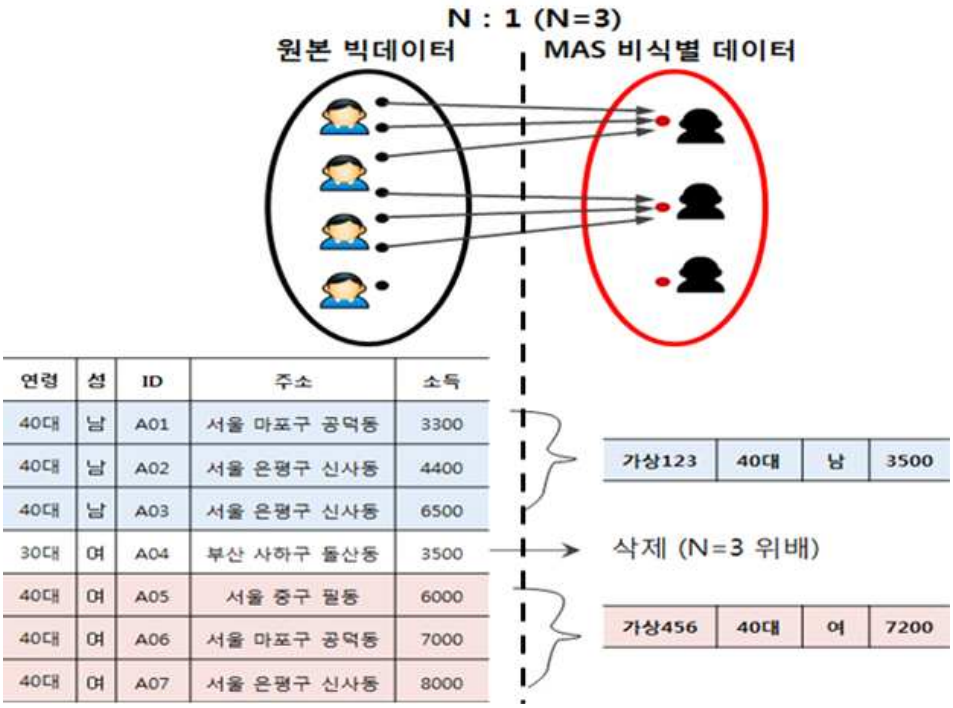
○ N명의 실제 개인을 1명의 추상화된 개인으로 변환(N:1)하여 다양한 빅데이터 분석이 가능한 나노(nano) 통계 빅데이터 생산 가능

나. 주요 알고리즘 특징

(1) MAS 비식별화 기술 용어 정리

기술 용어	설명
다수준	데이터 보안 및 활용 수준에 따라 저수준 및 고수준의 추상화/동기화 기술을 적용
추상화	원본 레코드의 그룹핑을 기반으로 하는 비식별화 기술
동기화	추상화된 비식별 데이터 세트 간의 연계 기술
추상화 키	레코드를 그룹핑 하기위해 기준이 되는 명목형 속성
추상화 속성	추상화 키 속성을 제외한 속성들 중 분석에 필요한 추상화 비식별 결과 레코드에 포함될 속성
추상화 크기(N)	추상화 키가 동일한 레코드를 그룹핑 할 때 N개의 레코드가 1개의 비식별 결과 레코드가 됨
동기화 연결 속성	각 테이블에 속성들 중 동일한 의미를 가진 속성을 뜻하며 동기화 수행시 의미적으로 공통 속성을 이용하여 연계 수행

(2) MAS 저수준 추상화

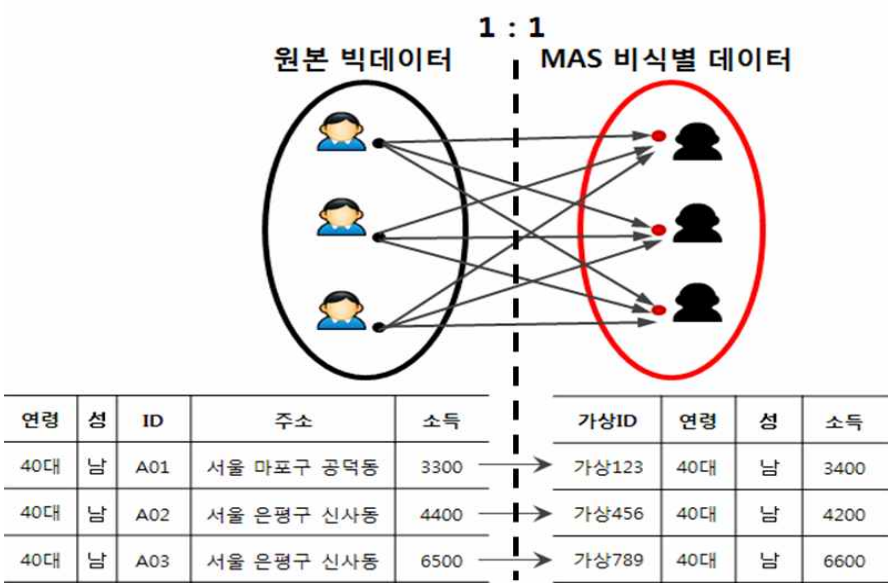


○ N명의 실제 개인을 1명의 추상화된 개인으로 변환 (N : 1)

○ 외부(인터넷) 공개 유통용 빅데이터 생산

○ 다양한 빅데이터 분석이 가능한 나노(nano)통계 빅데이터 생산 가능

(3) MAS 고수준 추상화



연령	성	ID	주소	소득	가상ID	연령	성	소득
40대	남	A01	서울 마포구 공덕동	3300	가상123	40대	남	3400
40대	남	A02	서울 은평구 신사동	4400	가상456	40대	남	4200
40대	남	A03	서울 은평구 신사동	6500	가상789	40대	남	6600

○ N명의 실제 개인을 N명의 추상화된 개인으로 변환 (1 : 1)

○ 외부 비공개 유통용 빅데이터 생산

○ 원본 빅데이터와의 유사성 극대화 (추상화 오류 보정 / 최적화 수행)

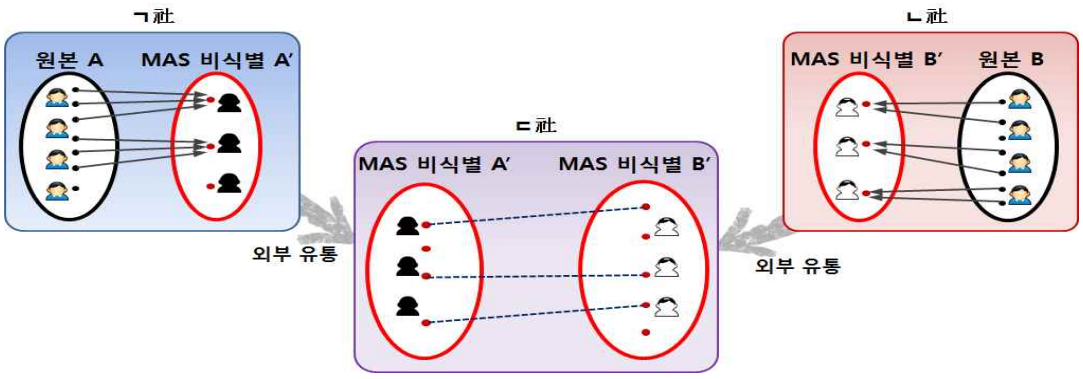
(4) 비식별화 결과 검증 지표

검증 지표	설명
잔존율	MAS 비식별화 과정에서 삭제되는 데이터 비율 및 잔존 비율

○ 저수준 추상화 : $\text{잔존율} = \frac{(\text{비식별결과레코드수} \times \text{추상화크기}(N))}{\text{원본레코드수}} \times 100$

○ 고수준 추상화 : $\text{잔존율} = \frac{\text{비식별결과레코드수}}{\text{원본레코드수}} \times 100$

(5) MAS 저수준 동기화



○ 비식별 변환된 데이터를 다른 비식별 데이터 또는 외부 빅데이터와 연계하여 다양한 분야의 빅데이터 유통 및 메쉬업, 분석 활용을 지원

○ 다양한 사용자 조건에 맞추어 공통 속성 기반의 연계를 수행함

ex) 〈가〉 테이블의 A속성과 〈나〉 테이블의 B 속성이 의미적으로 공통 속성인 경우, A와 B 속성값의 차이가 사용자 지정 조건 값보다 작으면 해당 레코드들은 결합이 이루어짐

○ 표준코드 필요 없이, 데이터 테이블 간의 공통 속성이 존재하면 자유롭게 연계가 가능한 데이터 연계 유연성을 확보할 수 있음

○ 개인 임시대체키 생성 없이 데이터 연계 가능

○ 연계를 수행하기 위한 전문기관 불필요함

(6) MAS 고수준 동기화

○ 비식별 변환된 데이터들을 연계하되, 저수준 동기화 방법보다 정확하게 결합을 수행하여 분석의 품질을 고도화

○ 2개 원본 데이터 간의 결합 결과와의 유사성 및 정확도를 극대화함

다. 데이터 비식별화 과정 상세

(1) 미수납이력(신용도) 데이터 추상화

○ 데이터 스키마

순번	미수납이력(신용도)				비고
	Field명	내용	데이터 타입	길이	
1	기준월		VARCHAR	8	
2	가입자 이름		VARCHAR	80	준식별자
3	이동전화번호		VARCHAR	20	준식별자
4	월 평균 통화시간	구간화	NUMERIC	15	
5	월 평균 통화빈도	구간화	INTEGER	8	
6	서비스 개월 수		NUMERIC	10	우선삭제대상
7	가입자 생년월		VARCHAR	8	YYYYMM
8	가입자 성별		VARCHAR	1	1(남), 2(여)
9	멤버쉽 등급		VARCHAR	1	
10	멤버쉽 카드 발급 여부		VARCHAR	1	우선삭제대상
11	거주지역-시도		VARCHAR	40	
12	거주지역-시군구		VARCHAR	80	
13	거주지역-읍면동		VARCHAR	200	
14	2nd Device 가입여부		VARCHAR	1	'Y', 'N'

순번	미수납이력(신용도)				비고
	Field명	내용	데이터 타입	길이	
15	ARPU	구간화	INTEGER	8	의미확인 필요
16	당월 납부요금	구간화	NUMERIC	18	
17	결합상품 가입여부		VARCHAR	1	삭제우선대상(2순위)
18	단말기 제조사		VARCHAR	80	삭제우선대상(1순위)
19	단말기종		VARCHAR	5	삭제우선대상(1순위)
20	단말기 출고가	구간화	NUMERIC	10	
21	가입일자	구간화	VARCHAR	8	
22	이용정지기간	구간화	INTEGER	8	
23	미납구분코드		VARCHAR	1	우선삭제대상, 의미확인 필요
24	당월연체유무		VARCHAR	1	
25	당월연체금액	구간화	NUMERIC	18	
26	최근 1년간 납부일 미준수 횟수	숫자	INTEGER	8	
27	최근 1년간 최대 연체금액	구간화	NUMERIC	18	
28	납부방법		VARCHAR	20	
29	회선상태		VARCHAR	20	
30	남은 할부 원금	구간화	NUMERIC	15	
31	남은 할부 잔여 기간	구간화	NUMERIC	8	

○ 데이터 전처리

- 원본 8,000,000의 데이터 중 가입자 생년월 속성에서 날짜형식(YYYYMM)에 부합하지 않은 레코드 1건 제거하였음
- 원본 속성에 대해 부분 마스킹, 이산화, 범주화를 수행하여 새로운 속성 생성 및 변환

원본 속성명	변환 속성명	설명
-	일련번호	MAS 비식별화에 필요한 PK속성이 존재하지 않기 때문에 레코드의 일련번호 속성을 생성
가입자 이름	가입자 성	가입자 이름 속성의 첫 글자를 추출하여 가입자 성 속성 생성
가입자 생년월일	연령	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령 속성 생성 : 2018 - 태어난 년도
	연령대	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령대 속성 생성 (2018 - 태어난 년도)/10하고 소수점 버림
가입일자	가입년도	가입일자 속성의 앞 4글자를 추출하여 가입년도 속성 생성
	가입월	가입일자 속성의 가운데 2글자를 추출하여 가입월 속성 생성
	가입일	가입일자 속성의 뒤 2글자를 추출하여 가입일 속성 생성

○ 추상화 파라미터 설정

- PK 속성 : 레코드별 일련번호를 생성하여 PK 속성으로 지정하였음
- 제거 속성 : 8개(PK, 멤버쉽 카드 발급 여부, 미납구분코드, 이동전화번호, 가입자 생년월, 가입자 이름, 단말기 종, 단말기 제조사)
- 추상화키 : 16개(기준월, 당월연체유무, 결합상품가입여부, 2nd Device 가입여부, 가입자 성별, 회선상태, 멤버쉽 등급, 납부방법, 연령대, 가입월, 거주지역-시도, 가입일, 가입년도, 거주지역-시군구, 가입자 성, 거주지역-읍면동)
- 추상화 속성 : 13개(최근 1년간 납부일 미준수 횟수, 남은 할부 잔여 기간, 연령, 서비스 개월 수, 단말기 출고가, 월 평균 통화빈도, 이용정지기간, 월 평균 통화시간, 당월연체금액, 최근 1년간 최대 연체금액, 당월 납부요금, 남은할부원금, ARPU)
- 추상화 레벨 : 고수준
- 추상화 크기(N) : 2
- k-익명성 검증 대상 속성 : 가입자 성, 가입자 연령대, 가입자 성별, 거주지역_시도, 거주지역_시군구, 거주지역_읍면동

○ 추상화 결과 스키마

미수납이력(신용도)- 비식별(추상화) 결과 데이터 스키마					
순번	한글명	순번	한글명	순번	한글명
1	기준월	12	가입일	23	당월납부요금
2	가입자_성	13	당월연체유무	24	단말기출고가
3	가입자성별	14	납부방법	25	이용정지기간
4	멤버쉽등급	15	회선상태	26	당월연체금액
5	거주지역_시도	16	ABST_ID	27	최근1년간납부일미준수횟수
6	거주지역_시군구	17	월평균통화시간	28	최근1년간최대연체금액
7	거주지역_읍면동	18	월평균통화빈도	29	남은할부원금
8	DOUBLE_DEVICE_여부	19	서비스개월수	30	남은할부잔여기간
9	결합상품가입여부	20	연령대		
10	가입년도	21	연령		
11	가입월	22	ARPU		

- 명목형 속성 수 : 16개(1~16)
- 수치형 속성 수 : 14개(17~30)

○ 추상화 결과

	k-익명성=4	k-익명성=6
원본 레코드 수	7,999,999 건	7,999,999 건
결과 레코드 수	7,898,566 건	7,788,927 건
잔존율	98.73%	97.36%

(2) 장애우 가입자 데이터 추상화

○ 데이터 스키마

순번	장애우 가입자				비고
	Field명	내용	데이터 타입	길이	
1	기준월		VARCHAR	8	
2	가입자이름		VARCHAR	80	추가(준식별자)
3	이동전화번호		VARCHAR	80	준식별자
4	가입자 생년월	YYYYMM	VARCHAR	20	
5	가입자 성별		VARCHAR	6	
6	장애인타입	시각,청각,일반,지체	VARCHAR	1	A,B,C,D
7	주중_06_09_방문지역-시도		VARCHAR	40	
8	주중_06_09_방문지역-시군구		VARCHAR	80	
9	주중_06_09_방문지역-읍면동		VARCHAR	200	
10	시간대1	06:00-09:00	VARCHAR	12	"06:00-09:00"로 세팅
11	주중_09_12_방문지역-시도		VARCHAR	40	
12	주중_09_12_방문지역-시군구		VARCHAR	80	
13	주중_09_12_방문지역-읍면동		VARCHAR	200	
14	시간대2	09:00-12:00	VARCHAR	12	"09:00-12:00"로 세팅
15	주중_12_14_방문지역-시도		VARCHAR	40	
16	주중_12_14_방문지역-시군구		VARCHAR	80	
17	주중_12_14_방문지역-읍면동		VARCHAR	200	
18	시간대3	12:00-14:00	VARCHAR	12	"12:00-14:00"로 세팅
19	주중_14_18_방문지역-시도		VARCHAR	40	
20	주중_14_18_방문지역-시군구		VARCHAR	80	
21	주중_14_18_방문지역-읍면동		VARCHAR	200	
22	시간대4	14:00-18:00	VARCHAR	12	"14:00-18:00"로 세팅
23	주중_18_22_방문지역-시도		VARCHAR	40	
24	주중_18_22_방문지역-시군구		VARCHAR	80	
25	주중_18_22_방문지역-읍면동		VARCHAR	200	
26	시간대5	18:00-22:00	VARCHAR	12	"18:00-22:00"로 세팅
27	주중_22_06_방문지역-시도		VARCHAR	40	
28	주중_22_06_방문지역-시군구		VARCHAR	80	
29	주중_22_06_방문지역-읍면동		VARCHAR	200	
30	시간대6	22:00-06:00	VARCHAR	12	"22:00-06:00"로 세팅
31	주말_06_09_방문지역-시도		VARCHAR	40	
32	주말_06_09_방문지역-시군구		VARCHAR	80	
33	주말_06_09_방문지역-읍면동		VARCHAR	200	
34	시간대7	06:00-09:00	VARCHAR	12	"06:00-09:00"로 세팅

순번	장애우 가입자				비고
	Field명	내용	데이터 타입	길이	
35	주말_09_12_방문지역-시도		VARCHAR	40	
36	주말_09_12_방문지역-시군구		VARCHAR	80	
37	주말_09_12_방문지역-읍면동		VARCHAR	200	
38	시간대8	09:00-12:00	VARCHAR	12	"09:00-12:00"로 세팅
39	주말_12_14_방문지역-시도		VARCHAR	40	
40	주말_12_14_방문지역-시군구		VARCHAR	80	
41	주말_12_14_방문지역-읍면동		VARCHAR	200	
42	시간대9	12:00-14:00	VARCHAR	12	"12:00-14:00"로 세팅
43	주말_14_18_방문지역-시도		VARCHAR	40	
44	주말_14_18_방문지역-시군구		VARCHAR	80	
45	주말_14_18_방문지역-읍면동		VARCHAR	200	
46	시간대10	14:00-18:00	VARCHAR	12	"14:00-18:00"로 세팅
47	주말_18_22_방문지역-시도		VARCHAR	40	
48	주말_18_22_방문지역-시군구		VARCHAR	80	
49	주말_18_22_방문지역-읍면동		VARCHAR	200	
50	시간대11	18:00-22:00	VARCHAR	12	"18:00-22:00"로 세팅
51	주말_22_06_방문지역-시도		VARCHAR	40	
52	주말_22_06_방문지역-시군구		VARCHAR	80	
53	주말_22_06_방문지역-읍면동		VARCHAR	200	
54	시간대12	22:00-06:00	VARCHAR	12	"22:00-06:00"로 세팅

○ 데이터 전처리

- 원본 885,433건의 데이터 중 가입자 생년월 속성에서 날짜형식(YYYYMM)에 부합하지 않은 레코드 80,352건 제거하였음
- 원본 속성에 대해 부분 마스킹, 이산화, 범주화를 수행하여 새로운 속성 생성 및 변환

원본 속성명	변환 속성명	설명
-	일련번호	MAS 비식별화에 필요한 PK속성이 존재하지 않기 때문에 레코드의 일련번호 속성을 생성
가입자 이름	가입자 성	가입자 이름 속성의 첫 글자를 추출하여 가입자 성 속성 생성
가입자 생년월일	연령	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령 속성 생성 : 2018 - 태어난 년도
	연령대	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령대 속성 생성(2018 - 태어난 년도)/10하고 소수점 버림

○ 추상화 파라미터 설정

- PK 속성 : 레코드별 일련번호 생성하여 PK 속성으로 지정하였음
- 제거 속성 : 5개(PK, 기준월, 가입자 이름, 이동전화번호, 가입자 생년월)
- 추상화키 : 52개(가입자 성별, 장애인 타입, 연령대. 가입자 성, 각 시간대별 속성 48개)
- 추상화 속성 : 1개(연령)
- 추상화 레벨 : 고수준
- 추상화 크기(N) : 2
- k-익명성 검증 대상 속성 : 가입자 성, 가입자 연령대, 가입자 성별, 장애인 타입

○ 추상화 결과 스키마

장애우 가입자 - 비식별(추상화) 결과 데이터 스키마					
순번	한글명	순번	한글명	순번	한글명
1	연령대	19	주중_14_18_방문지역-시군구	37	시간대8
2	연령	20	주중_14_18_방문지역-읍면동	38	주말_12_14_방문지역-시도
3	가입자 성	21	시간대4	39	주말_12_14_방문지역-시군구
4	가입자 성별	22	주중_18_22_방문지역-시도	40	주말_12_14_방문지역-읍면동
5	장애인타입	23	주중_18_22_방문지역-시군구	41	시간대9
6	주중_06_09_방문지역-시도	24	주중_18_22_방문지역-읍면동	42	주말_14_18_방문지역-시도
7	주중_06_09_방문지역-시군구	25	시간대5	43	주말_14_18_방문지역-시군구
8	주중_06_09_방문지역-읍면동	26	주중_22_06_방문지역-시도	44	주말_14_18_방문지역-읍면동
9	시간대	27	주중_22_06_방문지역-시군구	45	시간대10
10	주중_09_12_방문지역-시도	28	주중_22_06_방문지역-읍면동	46	주말_18_22_방문지역-시도
11	주중_09_12_방문지역-시군구	29	시간대6	47	주말_18_22_방문지역-시군구
12	주중_09_12_방문지역-읍면동	30	주말_06_09_방문지역-시도	48	주말_18_22_방문지역-읍면동
13	시간대2	31	주말_06_09_방문지역-시군구	49	시간대11
14	주중_12_14_방문지역-시도	32	주말_06_09_방문지역-읍면동	50	주말_22_06_방문지역-시도
15	주중_12_14_방문지역-시군구	33	시간대7	51	주말_22_06_방문지역-시군구
16	주중_12_14_방문지역-읍면동	34	주말_09_12_방문지역-시도	52	주말_22_06_방문지역-읍면동
17	시간대3	35	주말_09_12_방문지역-시군구	53	시간대12
18	주중_14_18_방문지역-시도	36	주말_09_12_방문지역-읍면동	54	ABST_ID

- 명목형 속성 수 : 52개(3～54)
- 수치형 속성 수 : 2개(1～2)

○ 추상화 결과

	k-익명성=4	k-익명성=6
원본 레코드 수	805,081건	805,081건
결과 레코드 수	779,979건	745,729건
잔존율	96.88%	92.62%

(3) 외국인 가입자 데이터 추상화

○ 데이터 스키마

순번	외국인 가입자				비고
	Field명	내용	데이터 타입	길이	
1	기준월		VARCHAR	8	
2	가입자이름		VARCHAR	80	추가(준식별자)
3	이동전화번호		VARCHAR	80	준식별자
4	가입자 생년월	YYYYMM	VARCHAR	20	
5	가입자 성별		VARCHAR	6	
6	장애인타입	시각,청각,일반,지체	VARCHAR	1	A,B,C,D
7	주중_06_09_방문지역-시도		VARCHAR	40	
8	주중_06_09_방문지역-시군구		VARCHAR	80	
9	주중_06_09_방문지역-읍면동		VARCHAR	200	
10	시간대1	06:00-09:00	VARCHAR	12	"06:00-09:00"로 세팅
11	주중_09_12_방문지역-시도		VARCHAR	40	
12	주중_09_12_방문지역-시군구		VARCHAR	80	
13	주중_09_12_방문지역-읍면동		VARCHAR	200	
14	시간대2	09:00-12:00	VARCHAR	12	"09:00-12:00"로 세팅
15	주중_12_14_방문지역-시도		VARCHAR	40	
16	주중_12_14_방문지역-시군구		VARCHAR	80	
17	주중_12_14_방문지역-읍면동		VARCHAR	200	
18	시간대3	12:00-14:00	VARCHAR	12	"12:00-14:00"로 세팅
19	주중_14_18_방문지역-시도		VARCHAR	40	
20	주중_14_18_방문지역-시군구		VARCHAR	80	
21	주중_14_18_방문지역-읍면동		VARCHAR	200	
22	시간대4	14:00-18:00	VARCHAR	12	"14:00-18:00"로 세팅
23	주중_18_22_방문지역-시도		VARCHAR	40	
24	주중_18_22_방문지역-시군구		VARCHAR	80	
25	주중_18_22_방문지역-읍면동		VARCHAR	200	
26	시간대5	18:00-22:00	VARCHAR	12	"18:00-22:00"로 세팅
27	주중_22_06_방문지역-시도		VARCHAR	40	
28	주중_22_06_방문지역-시군구		VARCHAR	80	
29	주중_22_06_방문지역-읍면동		VARCHAR	200	
30	시간대6	22:00-06:00	VARCHAR	12	"22:00-06:00"로 세팅
31	주말_06_09_방문지역-시도		VARCHAR	40	
32	주말_06_09_방문지역-시군구		VARCHAR	80	
33	주말_06_09_방문지역-읍면동		VARCHAR	200	

순번	외국인 가입자				비고
	Field명	내용	데이터 타입	길이	
34	시간대7	06:00-09:00	VARCHAR	12	"06:00-09:00"로 세팅
35	주말_09_12_방문지역-시도		VARCHAR	40	
36	주말_09_12_방문지역-시군구		VARCHAR	80	
37	주말_09_12_방문지역-읍면동		VARCHAR	200	
38	시간대8	09:00-12:00	VARCHAR	12	"09:00-12:00"로 세팅
39	주말_12_14_방문지역-시도		VARCHAR	40	
40	주말_12_14_방문지역-시군구		VARCHAR	80	
41	주말_12_14_방문지역-읍면동		VARCHAR	200	
42	시간대9	12:00-14:00	VARCHAR	12	"12:00-14:00"로 세팅
43	주말_14_18_방문지역-시도		VARCHAR	40	
44	주말_14_18_방문지역-시군구		VARCHAR	80	
45	주말_14_18_방문지역-읍면동		VARCHAR	200	
46	시간대10	14:00-18:00	VARCHAR	12	"14:00-18:00"로 세팅
47	주말_18_22_방문지역-시도		VARCHAR	40	
48	주말_18_22_방문지역-시군구		VARCHAR	80	
49	주말_18_22_방문지역-읍면동		VARCHAR	200	
50	시간대11	18:00-22:00	VARCHAR	12	"18:00-22:00"로 세팅
51	주말_22_06_방문지역-시도		VARCHAR	40	
52	주말_22_06_방문지역-시군구		VARCHAR	80	
53	주말_22_06_방문지역-읍면동		VARCHAR	200	
54	시간대12	22:00-06:00	VARCHAR	12	"22:00-06:00"로 세팅

○ 데이터 전처리

- 원본 319,110건의 데이터 중 가입자 생년월 속성에서 날짜형식(YYYYMM)에 부합하지 않은 레코드 7,198건 제거하였음
- 원본 속성에 대해 부분 마스킹, 이산화, 범주화를 수행하여 새로운 속성 생성 및 변환

원본 속성명	추가 속성명	설명
-	일련번호	MAS 비식별화에 필요한 PK속성이 존재하지 않기 때문에 레코드의 일련번호 속성을 생성
가입자 이름	가입자 성	가입자 이름 속성의 첫 글자를 추출하여 가입자 성 속성 생성
가입자 생년월일	연령	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령 속성 생성 : 2018 - 태어난 년도
	연령대	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령대 속성 생성(2018 - 태어난 년도)/10하고 소수점 버림

○ 추상화 파라미터 설정

- PK 속성 : 레코드별 일련번호 생성하여 PK 속성으로 지정하였음
- 제거 속성 : 6개(PK, 기준월, 가입자 이름, 이동전화번호, 가입자 생년월, 장애인 타입)
- 추상화키 : 51개(가입자 성별, 연령대. 가입자 성, 각 시간대별 속성 48개)
- 추상화 속성 : 1개(연령)
- 추상화 레벨 : 고수준
- 추상화 크기(N) : 2
- k-익명성 검증 대상 속성 : 가입자 성, 가입자 연령대, 가입자 성별

○ 추상화 결과 스키마

외국인 가입자 - 비식별(추상화) 결과 데이터 스키마					
순번	한글명	순번	한글명	순번	한글명
1	연령대	19	주중_14_18_방문지역-읍면동	37	주말_12_14_방문지역-시도
2	연령	20	시간대4	38	주말_12_14_방문지역-시군구
3	가입자 성	21	주중_18_22_방문지역-시도	39	주말_12_14_방문지역-읍면동
4	가입자 성별	22	주중_18_22_방문지역-시군구	40	시간대9
5	주중_06_09_방문지역-시도	23	주중_18_22_방문지역-읍면동	41	주말_14_18_방문지역-시도
6	주중_06_09_방문지역-시군구	24	시간대5	42	주말_14_18_방문지역-시군구
7	주중_06_09_방문지역-읍면동	25	주중_22_06_방문지역-시도	43	주말_14_18_방문지역-읍면동
8	시간대	26	주중_22_06_방문지역-시군구	44	시간대10
9	주중_09_12_방문지역-시도	27	주중_22_06_방문지역-읍면동	45	주말_18_22_방문지역-시도
10	주중_09_12_방문지역-시군구	28	시간대6	46	주말_18_22_방문지역-시군구
11	주중_09_12_방문지역-읍면동	29	주말_06_09_방문지역-시도	47	주말_18_22_방문지역-읍면동
12	시간대2	30	주말_06_09_방문지역-시군구	48	시간대11
13	주중_12_14_방문지역-시도	31	주말_06_09_방문지역-읍면동	49	주말_22_06_방문지역-시도
14	주중_12_14_방문지역-시군구	32	시간대7	50	주말_22_06_방문지역-시군구
15	주중_12_14_방문지역-읍면동	33	주말_09_12_방문지역-시도	51	주말_22_06_방문지역-읍면동
16	시간대3	34	주말_09_12_방문지역-시군구	52	시간대12
17	주중_14_18_방문지역-시도	35	주말_09_12_방문지역-읍면동	53	ABST_ID
18	주중_14_18_방문지역-시군구	36	시간대8		

- 명목형 속성 수 : 51개(3～53)
- 수치형 속성 수 : 2개(1～2)

○ 추상화 결과

	k-익명성=4	k-익명성=6
원본 레코드 수	311,912건	311,912건
결과 레코드 수	307,032건	303,750건
잔존율	98.43%	97.38%

(4) S신용평가(외부) 데이터 및 SKT 통신 데이터 동기화

○ SKT 통신 데이터 원본 스키마

순번	Field명	기준월	데이터 타입	크기	비고
1	가입자 이름		VARCHAR	80	준식별자
2	이동전화번호		VARCHAR	20	준식별자
3	가입자 생년월일		VARCHAR	8	준식별자
4	가입자 성별		VARCHAR	1	준식별자
5	거주지역-시군구		VARCHAR	80	준식별자 서울지역 한정
6	월 평균 통화시간_2016-3	2016-03	NUMERIC	15	
7	월 평균 통화빈도_2016-3	2016-03	INTEGER	8	
8	멤버쉽 등급_2016-3	2016-03	VARCHAR	1	
9	Tablet 보유여부_2016-3	2016-03	VARCHAR	1	
10	Smartwatch 보유여부_2016-3	2016-03	VARCHAR	1	
11	ARPU_2016-3	2016-03	INTEGER	8	
12	당월 납부요금_2016-3	2016-03	NUMERIC	18	
13	결합상품 가입여부_2016-3	2016-03	VARCHAR	1	
14	단말기 출고가_2016-3	2016-03	NUMERIC	10	
15	서비스가입일자_2016-3	2016-03	VARCHAR	8	
16	정지일수_2016-3	2016-03	INTEGER	8	
17	당월연체유무_2016-3	2016-03	VARCHAR	1	
18	당월연체금액_2016-3	2016-03	NUMERIC	18	
19	최근1년간납부일미준수횟수_2016-3	2016-03	NUMERIC		
20	최근1년간최대연체금액_2016-3	2016-03	NUMERIC		
21	납부방법_2016-3	2016-03	VARCHAR	20	
22	회선상태_2016-3	2016-03	VARCHAR	20	
23	남은 할부 원금_2016-3	2016-03	NUMERIC	15	
24	남은 할부 잔여 기간_2016-3	2016-03	NUMERIC	8	
25	월 평균 통화시간_2016-9	2016-09	NUMERIC	15	
26	월 평균 통화빈도_2016-9	2016-09	INTEGER	8	
27	멤버쉽 등급_2016-9	2016-09	VARCHAR	1	
28	Tablet 보유여부_2016-9	2016-09	VARCHAR	1	
29	Smartwatch 보유여부_2016-9	2016-09	VARCHAR	1	
30	ARPU_2016-9	2016-09	INTEGER	8	
31	당월 납부요금_2016-9	2016-09	NUMERIC	18	
32	결합상품 가입여부_2016-9	2016-09	VARCHAR	1	
33	단말기 출고가_2016-9	2016-09	NUMERIC	10	
34	서비스가입일자_2016-9	2016-09	VARCHAR	8	
35	정지일수_2016-9	2016-09	INTEGER	8	
36	당월연체유무_2016-9	2016-09	VARCHAR	1	
37	당월연체금액_2016-9	2016-09	NUMERIC	18	
38	최근1년간납부일미준수횟수_2016-9	2016-09	NUMERIC		
39	최근1년간최대연체금액_2016-9	2016-09	NUMERIC		
40	납부방법_2016-9	2016-09	VARCHAR	20	
41	회선상태_2016-9	2016-09	VARCHAR	20	
42	남은 할부 원금_2016-9	2016-09	NUMERIC	15	
43	남은 할부 잔여 기간_2016-9	2016-09	NUMERIC	8	

○ SKT 통신 데이터 전처리

- 원본 3,909,347건의 데이터 중 가입자 생년월 속성에서 날짜형식(YYYYMM)에 부합하지 않은 레코드 7,198건 제거하였음
- 유효성 필터링 : 가입자 생년월일 속성의 윤년이 아닌 년도에 윤일이 포함된 레코드 4건 제거하였음

  ex) 19620229 -> 1962년은 윤년이 아니므로, 해당 레코드는 제거 대상임
- 서비스가입일자_2016-03, 서비스가입일자_2016-09 속성을 일수 변환하는 과정에서 속성값 중 문자(#)가 포함되어 발생하는 오류 해결하기 위해 서비스가입일자_2016-03, 서비스가입일자_2016-09 속성값 중 문자(#)을 기준일로 값 변경
- 원본 속성에 대해 부분 마스킹, 이산화, 범주화를 수행하여 새로운 속성 생성 및 변환

원본 속성명	추가 속성명	설명
가입자 이름	가입자 성	가입자 이름 속성의 첫 글자를 추출하여 가입자 성 속성 생성
가입자 생년월일	연령	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령 속성 생성 : 2018 - 태어난 년도
	연령대	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령대 속성 생성(2018 - 태어난 년도)/10하고 소수점 버림
	생년월일_일수변환	가입자 생년월일 속성을 일수로 변환하여 생년월일_일수변환 속성 생성 : 기준일(19000101)-가입자생년월일
서비스가입일자_2016-3	가입일자_일수변환_1603	서비스가입일자_2016-3 속성을 일수로 변환하여 가입일수변환_1603 속성 생성 : 기준일(19840101)-서비스가입일자_2016-3
서비스가입일자_2016-9	가입일자_일수변환_1609	서비스가입일자_2016-9 속성을 일수로 변환하여 가입일수변환_1609 속성 생성 : 기준일(19840101)-서비스가입일자_2016-9

○ SKT 통신 데이터 추상화 파라미터 설정

- PK 속성 : 이동전화번호
- 제거 속성 : 5개(가입자 이름, 이동전화번호, 가입자 생년월일, 3월/9월에 해당하는 서비스가입일자)
- 추상화키 : 20개(가입자 성, 가입자 성별, 연령대, 거주지역-시군구, 3월/9월에 해당하는 멤버쉽등급, Tablet 보유여부, Smartwatch 보유여부, 결합상품가입여부, 당월연체유무, 최근1년간납부일미준수횟수, 납부방법, 회선상태)
- 추상화 속성 : 24개(연령, 생년월일_일수변환, 3월/9월에 해당하는 월평균통화시간, 월평균통화빈도, ARPU, 당월납부요금, 단말기출고가, 가입일자_일수변환, 정지일수, 당월연체금액, 남은할부원금, 남은할부잔여기간, 최근1년간최대연체금액)
- 추상화 레벨 : 고수준
- 추상화 크기(N) : 2
- k-익명성 검증 대상 속성 : 가입자 성, 연령대, 가입자성별, 거주지역_시군구
- l-다양성 검증 대상 속성 : 당월연체금액_2016-3, 당월연체금액_2016-9

○ SKT 통신 데이터 추상화 결과 스키마

SKT 통신 데이터 - 비식별 결과 데이터 스키마					
순번	한글명	순번	한글명	순번	한글명
1	ABST_ID	16	당월연체유무_1609	31	최근1년간최대연체금액_1603
2	가입자 성	17	납부방법_1609	32	남은할부원금_1603
3	가입자성별	18	회선상태_1609	33	남은할부잔여기간_1603
4	거주지역_시군구	19	연령	34	월평균통화시간_1609
5	멤버쉽등급_1603	20	연령대	35	월평균통화빈도_1609
6	Tablet 보유여부_2016-3	21	가입자생년월일_일수	36	ARPU_1609
7	Smartwatch보유여부_2016-3	22	월평균통화시간_1603	37	당월납부요금_1609
8	결합상품가입여부_1603	23	월평균통화빈도_1603	38	단말기출고가_1609
9	당월연체유무_1603	24	ARPU_1603	39	서비스가입일자_1603_일수_1609
10	납부방법_1603	25	당월납부요금_1603	40	정지일수_1609
11	회선상태_1603	26	단말기출고가_1603	41	당월연체금액_1609
12	멤버쉽등급_1609	27	서비스가입일자_1603_일수	42	최근1년간납부일미준수횟수_1609
13	Tablet 보유여부_2016-9	28	정지일수_1603	43	최근1년간최대연체금액_1609
14	Smartwatch보유여부_2016-9	29	당월연체금액_1603	44	남은할부원금_1609
15	결합상품가입여부_1609	30	최근1년간납부일미준수횟수_1603	45	남은할부잔여기간_1609

- 명목형 속성 수 : 18개(1~18)
- 수치형 속성 수 : 27개(19~45)

○ SKT 통신 데이터 추상화 결과

	k-익명성=4 / l-다양성=2
원본 레코드 수	3,789,213건
결과 레코드 수	3,754,040건
잔존율	99.07%

○ S신용평가 원본 데이터 전처리

- 핸드폰번호 속성값이 NULL인 레코드 분리

	속성 수	총 레코드 수	비고
원본 데이터	570개	35,533,167 건	
핸드폰번호 속성값에 NULL이 포함된 레코드	570개	19,978,118 건	속성값이 NULL인 경우 고수준 동기화 수행이 어려움
정상적으로 핸드폰번호 속성값이 들어간 레코드	570개	15,555,049 건	고수준 동기화 대상 데이터 핸드폰번호 중복 5,783 건 존재

- 원본 속성에 대해 부분 마스킹, 이산화, 범주화를 수행하여 새로운 속성 생성 및 변환

원본 속성명	추가 속성명	설명
가입자 이름	가입자 성	가입자 이름 속성의 첫 글자를 추출하여 가입자 성 속성 생성
가입자 생년월일	연령	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령 속성 생성 : 2018 - 태어난 년도
	연령대	가입자 생년월일 속성에서 태어난 년도를 추출하여 아래 식으로 연령대 속성 생성(2018 - 태어난 년도)/10하고 소수점 버림

- 데이터 특이사항 : 데이터 추출 시점이 다른 경우 핸드폰번호와 이름이 중복되면서 가입자생년월일과 연령이 다른 레코드가 존재하였음

○ S신용평가(외부) 원본 데이터 및 SKT 추상화 데이터 고수준 동기화

- 동기화 방법
  - 동기화 연결 속성으로 가입자 성, 연령대, 가입자 성별로 지정하여 고수준 동기화 수행하였음
  - 임시대체키를 생성하지 않고 연계 수행
  - 전문기관의 개입 없이 MAS 동기화를 통하여 직접 연계 수행
- 동기화 결과

SKT 연계 대상 건수	S신용평가 연계 대상 건수	연계 성공 레코드 수
3,754,040 건	15,555,049 건	970,553 건

라. 과제를 통한 솔루션 고도화 내역

(1) 원본 스키마를 유지하면서 추상화 수행이 가능

○ 기존 MAS 비식별화(추상화) 알고리즘은 원본 스키마를 유지하기 어려움

○ 과제를 통해 원본 스키마를 유지하면서 재식별이 불가능하며 원본과의 통계적 유사성이 보존되는 비식별화 기술로 고도화됨

마. 외부기관 검증 수행

(1) TTA V&V(Verification & Validation) 인증 수행

○ 원본 유사도 검증
- 추상화 기반 비식별화로 변환된 레코드 세트와 원본 레코드 세트 간의 유사도가 90% 이상 보존되는지 확인
- 시료데이터는 국민건강보험에서 제공하는 공개 데이터 중 “건강 검진 정보” 데이터 사용

○ 원본 유사도 검증 결과
- 31가지 속성 값을 가진 100,000건의 데이터가 담겨있는 테이블에 대해 추상화 기반 비식별화를 수행했을 시, 원본 테이블과 변경된 테이블이 90%이상 유사한지 확인함
- 100,000건 데이터가 담긴 테이블 10개를 시료로 사용하여 시험하였고, 10개 테이블에 대하여 총 3회 시험 결과 원본테이블과 평균 98.64% 유사도가 유지되는 것을 확인함

5. 비식별화 비교 분석 (연세대)

가. k-익명성 기법/MAS 비식별 알고리즘 실증 변환 결과 분석

(1) k-익명성 기법/MAS 비식별 결과 데이터 비교 분석 개요

(가) 실증데이터: 신용도

<표 3-7> 신용도 원본, 비식별 데이터 각 레코드 수

구분	레코드 수	속성 수
원본	7,999,999	31
k-익명성 기법(K=3)	7,912,142	22
k-익명성 기법(K=5)	7,912,142	22
MAS(K=3)	7,898,566	26
MAS(K=5)	7,788,927	26

- 신용도 데이터 원본의 36개의 속성에 대하여 식별위험이 높은 속성 삭제(이동전화번호, 멤버쉽카드발급여부, 단말기제조사, 미납구분코드, 서비스개월수, 단말기종) 및 단순정보 속성(기준월) 삭제
- k-익명성 기법/MAS 비식별 조치 결과 데이터의 속성 수 차이는 k-익명성 기법의 경우 거주지역-시군구, 거주지역-읍면동 속성 삭제, MAS의 경우 가입일자를 가입년도, 가입월, 가입일 속성으로 분리
- 다음과 같은 공통 속성에 대하여 분석 수행

<표 3-8> 신용도 데이터 결과 분석 속성

속성 구분	대상 속성	속성 개수
준식별자 속성	고객명, 가입자 생년월, 가입자 성별, 거주지역,	6
명목형 민감속성	당월연체유무, 결합상품가입여부, Double Device 여부, 회선상태, 멤버쉽등급, 납부방법, 가입일자	7
수치형 민감속성	단말기출고가, 월평균통화빈도, 이용정지기간, 월평균통화시간, 당월연체금액, 최근1년간최대연체금액, 남은할부원금, 당월납부요금, ARPU, 최근1년간납부일미준수횟수, 남은할부잔여기간	11

(나) 실증데이터: 장애우 거소지

<표 3-9> 장애우 거소지 원본, 비식별 데이터 각 레코드 수

구분	레코드 수	속성 수
원본	805,081	54
k-익명성 기법(K=3)	805,080	40
k-익명성 기법(K=5)	805,081	40
MAS(K=3)	779,980	41
MAS(K=5)	745,730	41

- 장애우 거소지 데이터 원본의 54개의 속성에 대하여 식별위험이 높은 속성 삭제(이동전화번호) 및 단순정보 속성 삭제(기준월, 기준시간)
- k-익명성 기법/MAS 비식별 조치 결과 데이터의 속성 수 차이는 MAS의 경우 원본 데이터 가입자생년월 속성을 연령과 연령대 두 속성으로 비식별 조치
- 다음과 같은 공통 속성에 대하여 분석 수행

<표 3-10> 장애우 거소지 데이터 결과 분석 속성

속성 구분	대상 속성	속성 개수
준식별자 속성	가입자이름, 가입자생년월, 가입자성별, 장애인타입	4
명목형 민감속성	주중_06_09_방문지역_시도, 주중_06_09_방문지역_시군구, 주중_06_09_방문지역_읍면동, ... 주말_22_06_방문지역_읍면동	36

(다) 실증데이터: 외국인 체류지

<표 3-11> 외국인 체류지 원본, 비식별 데이터 각 레코드 수

구분	레코드 수	속성 수
원본	311,912	53
k-익명성 기법(K=3)	311,912	39
k-익명성 기법(K=5)	311,912	39
MAS(K=3)	307,033	40
MAS(K=5)	303,750	40

- 외국인 체류지 데이터 원본의 53개의 속성에 대하여 식별위험이 높은 속성 삭제(이동전화번호) 및 단순정보 속성 삭제(기준월, 기준시간)
- k-익명성 기법/MAS 비식별 조치 결과 데이터의 속성 수 차이는 MAS의 경우 원본 데이터 가입자생년월 속성을 연령과 연령대 두 속성으로 비식별 조치
- 다음과 같은 공통 속성에 대하여 분석 수행

<표 3-12> 외국인 체류지 데이터 결과 분석 속성

속성 구분	대상 속성	속성 개수
준식별자 속성	가입자이름, 가입자생년월, 가입자성별	3
명목형 민감속성	주중_06_09_방문지역_시도, 주중_06_09_방문지역_시군구, 주중_06_09_방문지역_읍면동, ... 주말_22_06_방문지역_읍면동	36

(2) k-익명성 기법/MAS 비식별 알고리즘 실증 변환 결과의 통계적 유사성 분석

(가) 통계적 유사성 분석의 필요성
- 원본 식별 데이터와 비식별 조치된 데이터 사이의 통계적 유사성을 정량적으로 평가함으로써 유통 전 데이터의 활용성 분석

(나) 분석 방법

① 원본유사도

원본ID	성별	수입	나이
A1	여	1500	23
A2	남	1700	32
A3	여	2900	43
A4	남	2100	25
A5	여	3000	40
A6	여	1900	28

결과ID	성별	수입	나이
X1	*	1532	20대
X2	*	1697	30대
X3	*	2835	40대
X4	*	2133	20대
X5	*	3013	40대
X6	*	1858	20대

[그림 3-13] 원본유사도 이해를 위한 비식별 조치 예시

- 원본 데이터와 비식별 조치된 데이터 간 통계적 유사성을 0과 1 사이의 지표로 표현
- 해당 데이터 레코드 셋에 대하여 각 속성별 비식별화 값에 대응되는 원본 값과의 통계적 유사 정도를 속성 유사도로 정의
- 속성별 원본유사도는 해당 속성의 타입(수치형, 명목형)에 따라 계산 방법이 상이함
- 수치형 속성의 경우 원본 도메인 대비 변화율을 원본 유사도로 정의

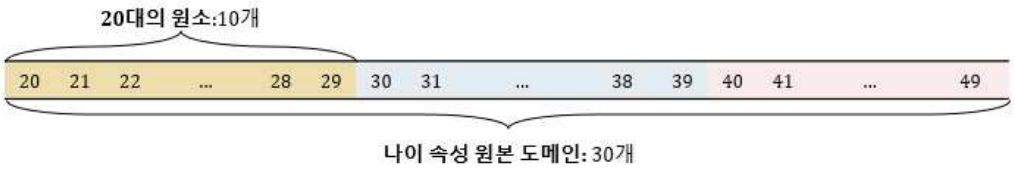
  예) $(\mathrm{A4}-\mathrm{X4}$쌍의 수입 속성 유사도$)$

  $= 1 - \frac{|2100-2133|}{Range(\text{수입})} = 1 - \frac{|2100-2133|}{\max(\text{수입}) - \min(\text{수입})} = 1 - \frac{|2100-2133}{3000-1400} = 1 - \frac{33}{1600} = 0.9793$

- 명목형 속성의 경우 원본 도메인의 고유값 개수 대비 비식별 결과에 해당하는 값 개수의 비율에 따라 원본 유사도를 정의

  예) $(\mathrm{A4}-\mathrm{X4}$쌍의 나이 속성 유사도$)$

  $= 1 - \frac{\text{비식별화 결과의 원소 } count \ - \ 1}{\text{원본도메인의 } distinct \ count} = 1 - \frac{10-1}{30} = 0.7$



[그림 3-14] 명목형 변환 속성 유사도 예시

- 해당 데이터 각 레코드에 존재하는 속성 유사도의 평균을 **레코드 유사도**로 정의

  예) $(\text{레코드 유사도}) = \frac{\sum(\text{속성 유사도})}{\text{속성 수}}$

  $= \frac{\text{성별 속성 유사도} + \text{수입 속성 유사도} + \text{나이 속성 유사도}}{3}$

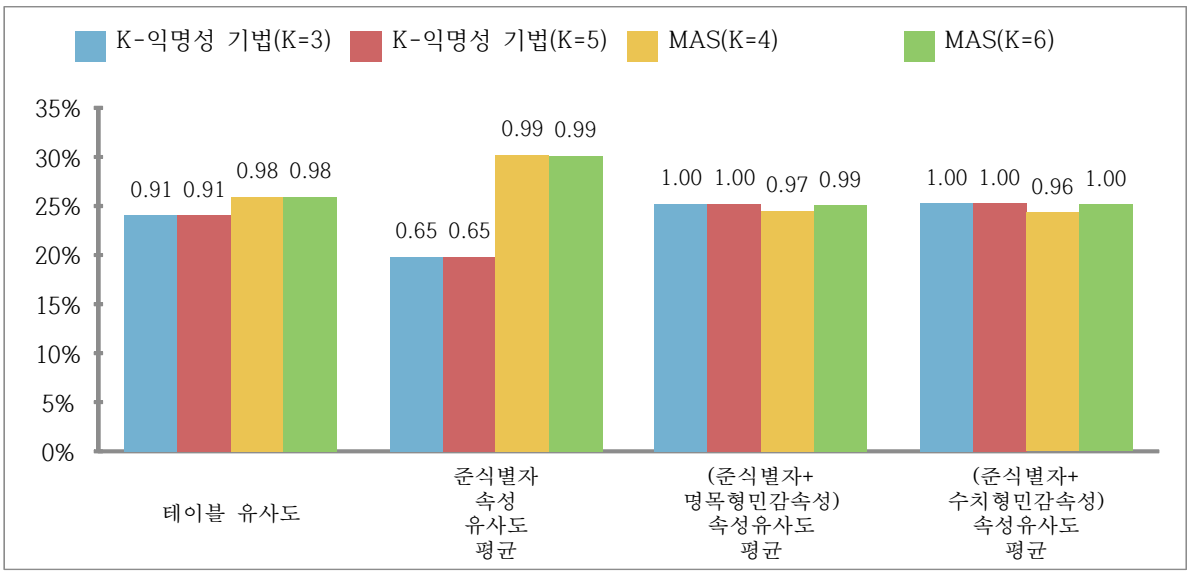
- 해당 데이터 레코드 셋의 모든 레코드 유사도의 평균을 **테이블 유사도**로 정의

② 잔존율

- 원본 데이터 셋의 레코드 수 대비 비식별 조치된 데이터 셋의 레코드 수를 백분율로 표기 및 분석
- 비식별 조치 과정에서 삭제되는 데이터의 비율을 지표로 분석함으로써 비식별 조치된 유통 데이터의 활용성 평가

(다) 신용도 실증데이터에 대한 분석 결과

① 각 기법별 원본 유사도

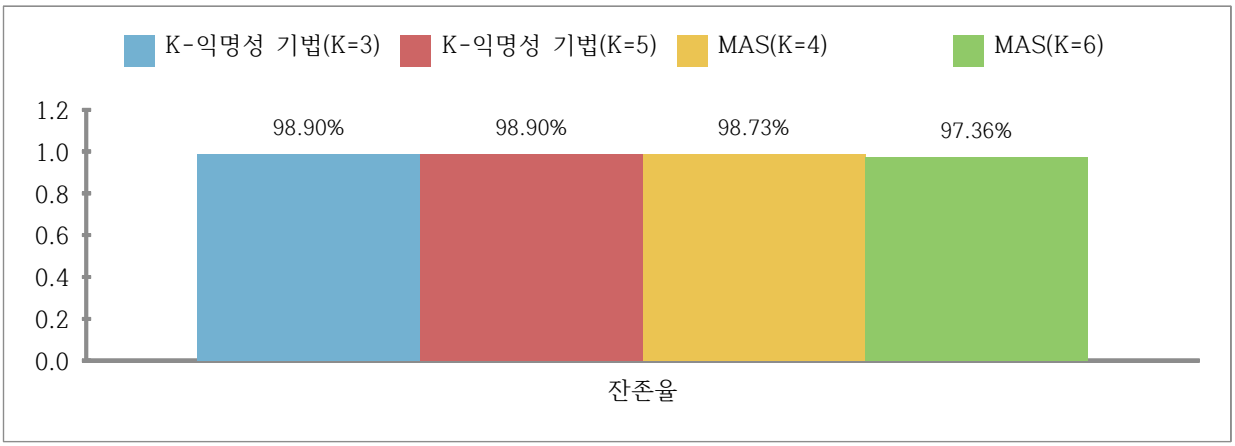


[그림 3-15] 비식별화 기법에 따른 원본 유사도 측정 결과: 신용도 데이터

- 전체 속성 조합의 테이블 유사도 분석 결과 두 기법 모두 테이블 유사도가 90% 이상으로 데이터 유통 시 분석 활용 가능할 것으로 예상되나 상대적으로 k-익명성 기법의 비식별화 결과가 유사성이 낮음

- 이는 준식별자 속성 중 거주지역 속성 중 시군구/읍면동 속성을 지워내면서 해당 속성의 유사도가 크게 떨어지며 발생한 현상

② 각 기법별 잔존율

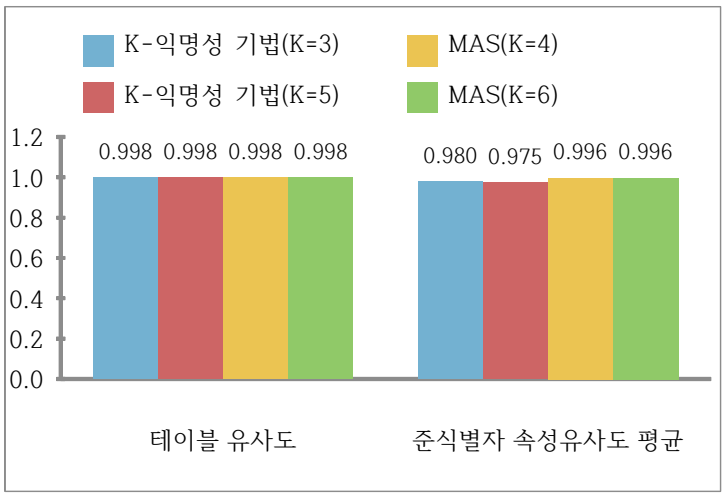


[그림 3-16] 비식별화 기법에 따른 잔존율 측정 결과 : 신용도 데이터

- 잔존율이 양 기법 모두 약 97% 이상으로 데이터 유통에 분석 활용 가능할 것으로 예상
- 일부 제거된 레코드는 재식별 가능 레코드로 유통이 완전 불가한 것으로 여겨짐

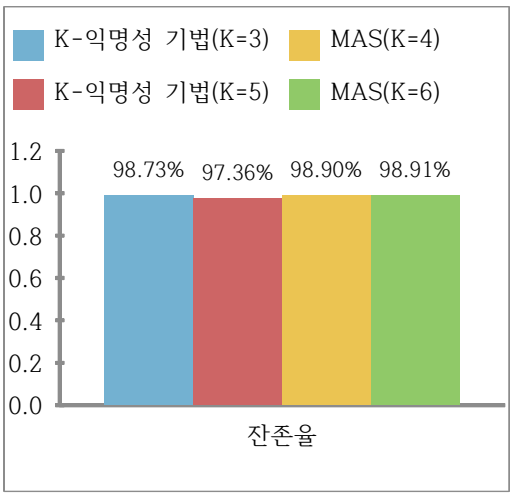
(라) 장애우 거소지 실증데이터에 대한 분석 결과

① 각 기법 별 테이블 유사도



[그림 3-17] 비식별화 기법에 따른 원본 유사도 측정 결과: 장애우 거소지

② 각 기법 별 잔존율

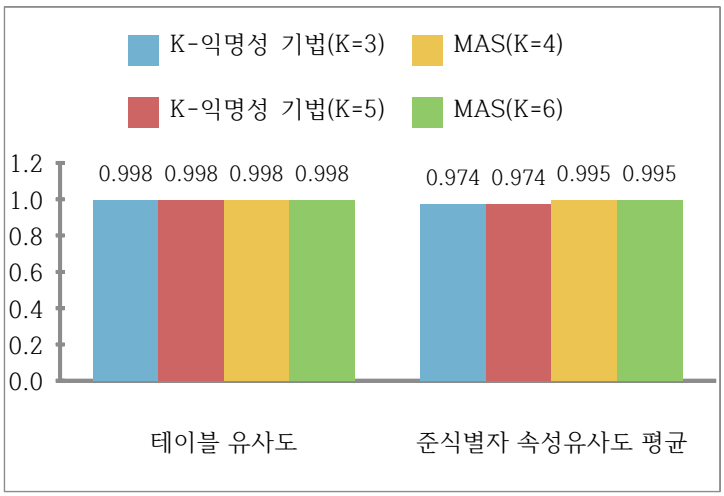


[그림 3-18] 비식별화 기법에 따른 잔존율 측정 결과: 장애우 거소지

- 전체 속성 조합의 테이블 유사도 분석 결과 두 기법 모두 유사도가 97% 이상으로 데이터 유통 시 분석 활용 가능할 것으로 예상됨
- 잔존율이 양 기법 모두 약 97% 이상으로 데이터 유통에 분석 활용 가능할 것으로 예상
- 일부 제거된 레코드는 재식별 가능 레코드로 유통이 완전 불가한 것으로 여겨짐

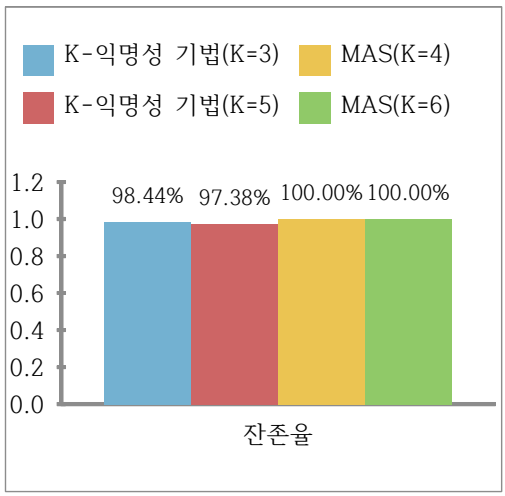
(마) 외국인 체류지 실증데이터 대한 분석 결과

① 각 기법별 테이블 유사도



[그림 3-19] 비식별화 기법에 따른 원본 유사도 측정 결과: 외국인 체류지

② 각 기법별 잔존율



[그림 3-20] 비식별화 기법에 따른 잔존율 측정 결과: 외국인 체류지

- 전체 속성 조합의 테이블 유사도 분석 결과 두 기법 모두 유사도가 97% 이상으로 데이터 유통 시 분석 활용 가능할 것으로 예상됨

- 잔존율이 양 기법 모두 약 97% 이상으로 데이터 유통에 분석 활용 가능할 것으로 예상
- 일부 제거된 레코드는 재식별 가능 레코드로 유통이 완전 불가한 것으로 여겨짐

(3) k-익명성 기법/MAS 비식별 알고리즘 실증 변환 결과의 재식별 가능성 분석

(가) 재식별 가능성 분석의 필요성

- 비식별 조치된 레코드에 대하여 준식별자를 포함한 민감 속성들의 연결공격(linkage attack) 가능성을 정량적으로 평가하여 데이터 유통 전 안전성을 분석
- 원본 데이터 세트 중 유일한 속성 값을 갖는 레코드가 비식별 데이터 세트에서도 동일 속성값을 가지면서 유일하게 계속 남아있다면 연결 공격에 의한 재식별 위험에 노출될 수 있으므로 이에 대한 재식별 가능성 분석이 필요

(나) 분석 방법

① m-유일성 (m-uniqueness) 검사

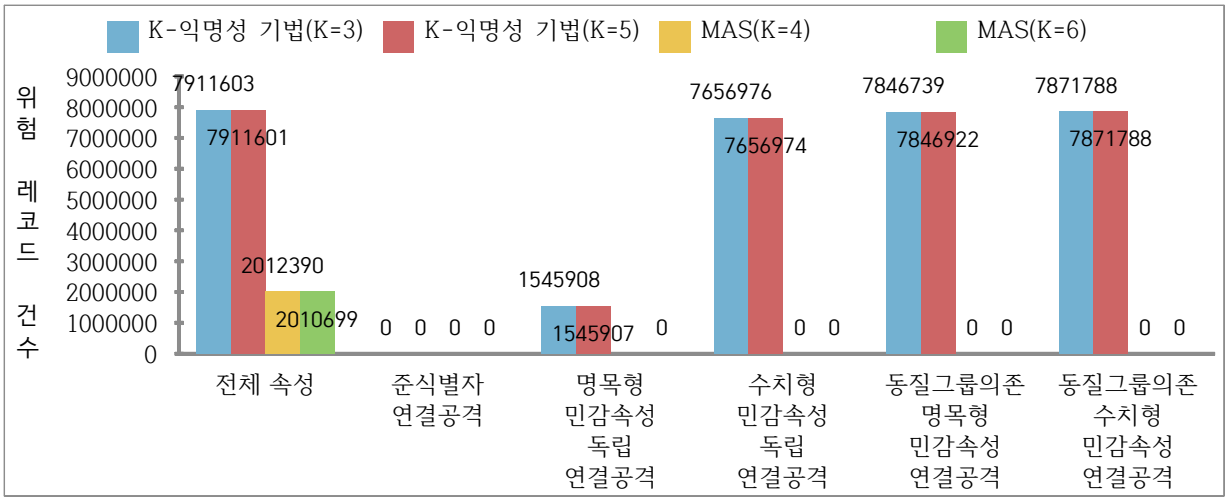
- 원본 데이터와 동일한 속성 값의 조합이 비식별 결과 데이터에 최소 m개 존재해야 재식별 가능성 위험이 낮음
- m-유일성 검사 방법은 속성 조합 별로 n-차원 검사를 수행
- 비식별 조치된 결과 데이터에 대하여 준식별자/민감속성 구분 및 속성 타입 구분에 따라 6 종류 공격 타입을 분류하고 그에 따른 속성 조합을 선정
- 각 속성 조합 별로 원본 식별 데이터 속성 내에서 유일 속성조합 값이 변환된 결과 데이터 레코드에서도 동일하게 유일한 속성 조합 값의 개수를 집계

<표 3-13> 준식별자/민감속성 구분 및 속성 타입 구분에 따른 속성 조합

	준식별자 속성	수치형 민감속성	명목형 민감속성
Type 1: 전체 속성	O	O	O
Type 2: 준식별자 재식별 위험성	O		
Type 3: 명목형 민감속성 독립 연결 공격 위험성		O	
Type 4: 수치형 민감속성 독립 연결 공격			O
Type 5: 동질그룹 의존 명목형 민감속성 연결 공격	O	O	
Type 6: 동질그룹 의존 수치형 민감속성 연결 공격	O		O

(다) 분석 결과

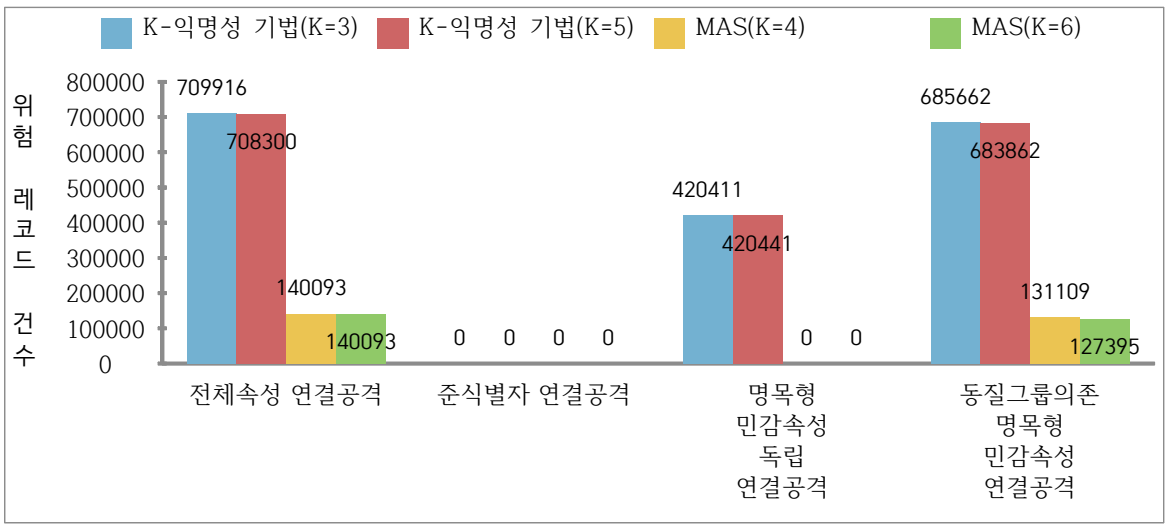
① 신용도 실증데이터에 대한 유일성 분석 결과



[그림 3-21] 비식별화 기법에 따른 m 유일성 측정 결과: 신용도

- k-익명성 기법 / MAS 기법은 모두 준식별자 속성 조합에 대한 재식별 위험성은 0으로 안전
- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합으로 연결공격시 취약함
- KLT 프라이버시 모델을 적용 시, k-익명성의 취약한 부분을 보강하여 위험성을 현저하게 낮출 수 있음
- MAS 기법의 경우 대부분 속성 조합의 재식별 위험 레코드의 수가 0건으로 k-익명성 기법와 비교하여 현저하게 낮은 위험성을 보유함

② 장애우 거소지 실증데이터에 대한 유일성 분석 결과

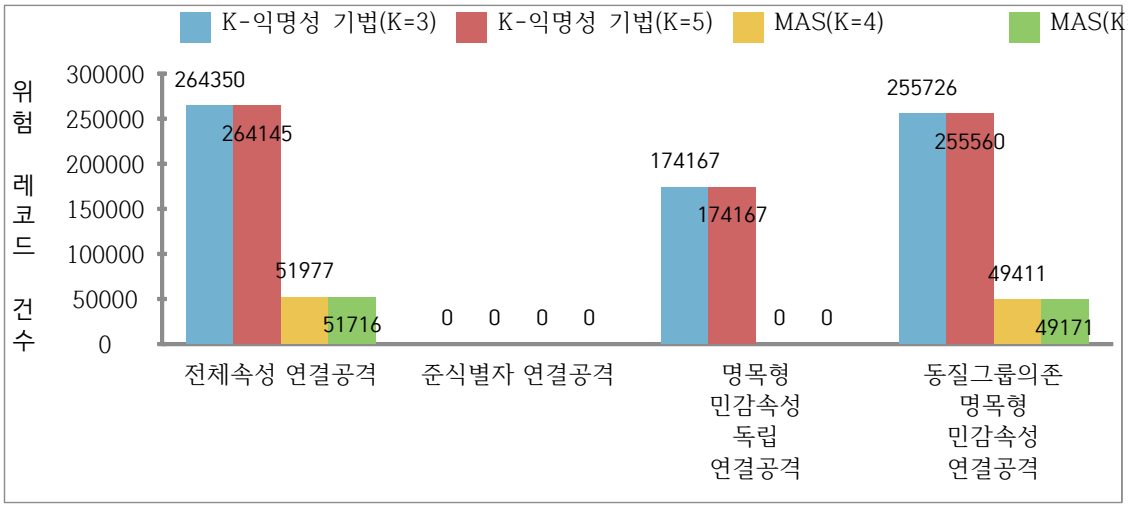


[그림 3-22] 비식별화 기법에 따른 m 유일성 측정 결과: 장애우 거소지

- k-익명성 기법 / MAS 기법은 모두 준식별자 속성 조합에 대한 재식별 위험성은 0으로 안전
- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합으로 연결공격시 취약함

- KLT프라이버시 모델을 적용시, k-익명성의 취약한 부분을 보강하여 위험성을 현저하게 낮출 수 있음
- MAS 기법의 경우 대부분 속성 조합의 재식별 위험 레코드의 수가 0건으로 k-익명성 기법와 비교하여 현저하게 낮은 위험성을 보유함

③ 외국인 체류지 실증데이터에 대한 유일성 분석 결과



[그림 3-23] 비식별화 기법에 따른 m-유일성 측정 결과: 외국인 체류지

- k-익명성 기법 / MAS 기법은 모두 준식별자 속성 조합에 대한 재식별 위험성은 0으로 안전
- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합으로 연결공격시 취약함
- KLT프라이버시 모델을 적용시, k-익명성의 취약한 부분을 보강하여 위험성을 현저하게 낮출 수 있음
- MAS 기법의 경우 대부분 속성 조합의 재식별 위험 레코드의 수가 0건으로 k-익명성 기법과 비교하여 현저하게 낮은 위험성을 보유함

나. 개인정보 비식별 데이터의 연계 활용성 평가

(1) 개인정보 비식별 데이터의 연계 데이터 요약

(가) 실증데이터: 신용도데이터 A, B

<표 3-14> 신용도 연계 원본, 비식별 데이터 각 레코드 수

구분	레코드 수(건)	속성 수(개)
원본 A	100,000	17
원본 B	100,000	17
k-익명성 기법 결과 A	100,000	14
k-익명성 기법 결과 B	109,328	14
k-익명성 기법 연계 결과	109,328	24
MAS 결과 A	89,994	16
MAS 결과 B	85,404	16
MAS 연계 결과	77,219	26

- 신용도 데이터 A 원본의 17개의 속성 중 식별위험이 높은 속성 삭제(이동전화번호)
- 신용도 데이터 B 원본의 17개의 속성 중 식별위험이 높은 속성 삭제(이동전화번호)
- k-익명성 기법/MAS 비식별 조치 결과 데이터의 속성 수 차이는 k-익명성 기법의 경우 거주지역-시군구, 거주지역-읍면동 속성 삭제
- 다음과 같은 공통 속성에 대하여 분석 진행

<표 3-15> 신용도 연계 데이터 결과 분석 속성

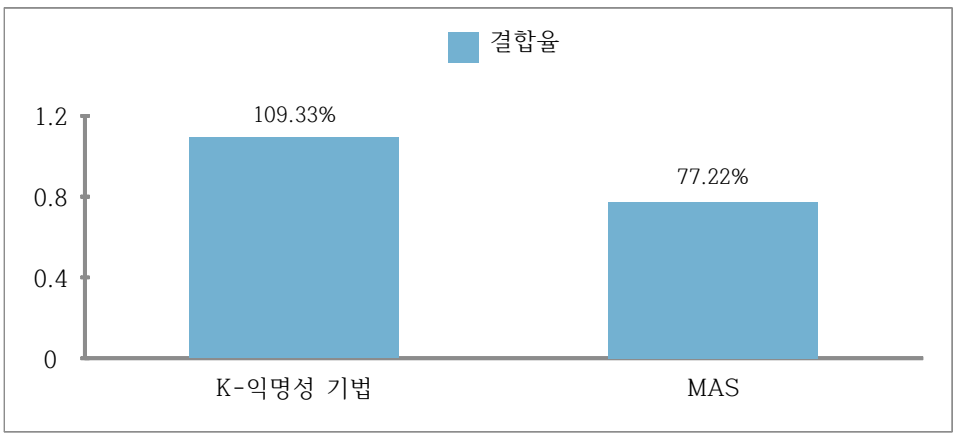
속성 구분	대상 속성	속성 개수
준식별자 속성	가입자명, 가입자생년월, 가입자성별, 거주지역-시도	4
명목형 민감속성	A.결합상품가입여부, A.가입일자, A.납부방법, B.멤버쉽등급, B.2ndDevice가입여부, B.당월연체유무, B.회선상태	7
수치형 민감속성	A.월평균통화시간, A.월평균통화빈도, A.ARPU, A.당월납부요금, A.단말기출고가, A.남은할부원금, A.남은할부잔여기간, B.멤버쉽사용금액, 당년멤버쉽사용금액, B.이용정지기간, B.당월연체금액, B.최근1년간납부일미준수횟수, 최근1년간최대연체금액	13

(2) 개인정보 비식별 데이터의 연계 데이터 결과의 통계적 유사성 분석

(가) 연계 결합율

- 원본 데이터의 연계 결과 데이터 세트의 레코드 수 대비 비식별 조치 된 데이터의 연계 결과 데이터 세트의 레코드 수의 비율을 결합율로 정의
- 연계 결합율은 연계 결과 데이터의 활용성 및 분석 결과의 정확성을 판단하는 기준이 됨

(나) 각 기법별 연계 결합율



[그림 3-24] 연계 기법에 따른 결합률 비교

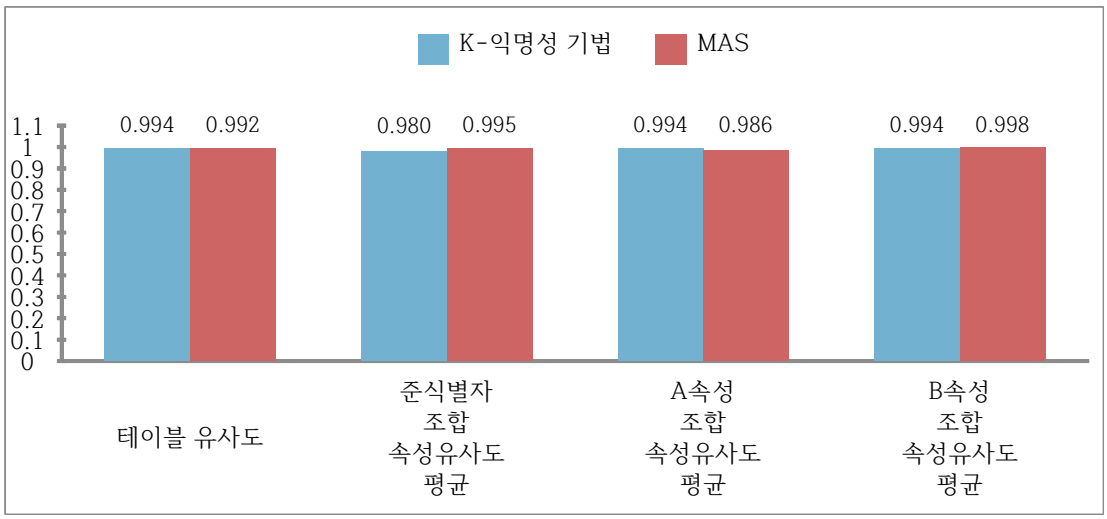
- k-익명성 기법 연계 방식의 경우 원본 대비 결합율이 100%을 넘어선 값을 보이며, 실제 예상 연계 레코드 수 10만 보다 많은 수의 레코드가 연계되는 현상을 보임
- 최소 9328건의 중복되는 불확실한 원본레코드 연계가 발생한 것으로 보여짐
- MAS 연계 방식의 경우 원본 대비 결합율이 0.77로 예상 연계 레코드 수 10만 건 보다 적은 수의 레코드가 연계되는 현상을 보임

- 이는 추상화 비식별화 단계에서 재식별 공격 위험성을 감소시키기 위해 손실된 레코드가 연계되지 못하여 발생한 문제로 보여짐

(다) 연계 데이터 원본유사성 분석

- 연계 데이터의 경우 A사 식별 원본 데이터와 B사 식별 원본 데이터를 결합한 원본 결합 데이터와 A사 비식별 조치된 데이터와 B사 비식별 조치된 데이터를 결합한 결과 연계 데이터의 통계적 유사성을 비교 분석함

(라) 각 기법별 테이블 유사도

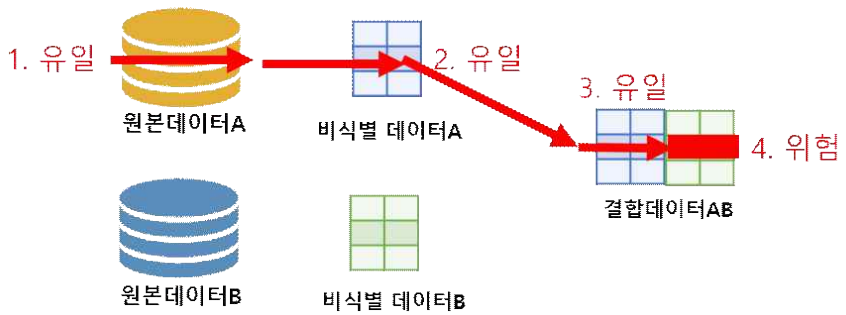


[그림 3-25] 연계 기법에 따른 원본 유사도 비교

- 테이블 유사도 분석 결과 두 기법 모두 데이터 변환율이 2% 이하로 데이터 유통 시 분석 활용도가 크게 저하되지 않은 것으로 판단됨

(4) 개인정보 비식별 데이터의 연계 데이터 결과의 재식별 가능성 분석

(가) 연계 데이터 재식별 가능성 분석

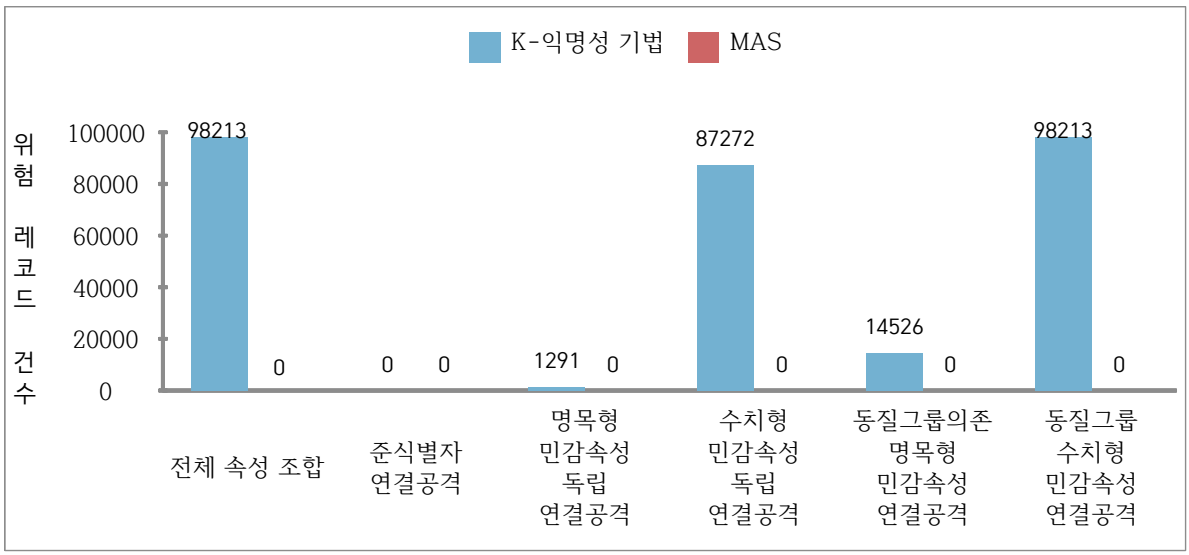


[그림 3-26] 연계 데이터 재식별 가능성 예제

예) 원본 데이터 A에서 유일했던 레코드가 비식별 데이터 A에서도 유일하다. 그 레코드가 결합데이터 AB에서도 유일했을 때 연결 공격(linkage attack)에 의해 결합되었던 B의 민감 속성이 식별될 위험이 생김

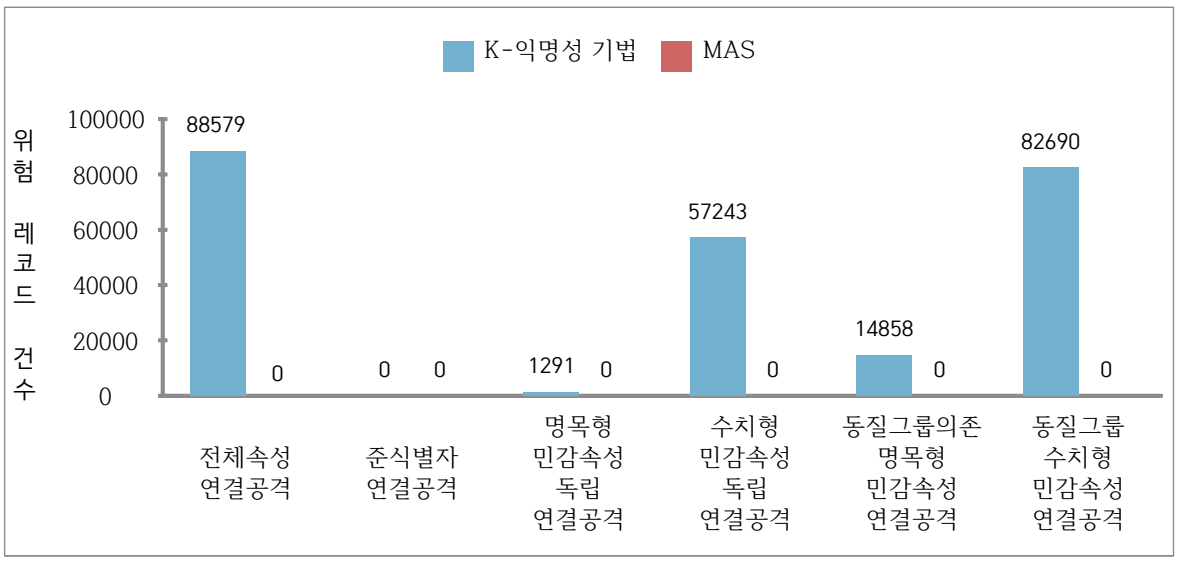
(나) 연계 데이터 재식별 가능성 분석에 대한 결과

① k-익명성 기법/MAS 결과 데이터 A에 대한 유일성 결과 검사 비교



[그림 3-27] 연계 기법에 따른 재식별 위험 레코드 건수 비교 : 신용도 A

② k-익명성 기법/MAS 결과 데이터 B에 대한 유일성 결과 검사 비교



[그림 3-28] 연계 기법에 따른 재식별 위험 레코드 건수 비교 : 신용도 B

- k-익명성 기법은 준식별자 속성 조합의 경우 재식별 가능성 위험이 없지만 빅데이터 활용을 위해 원본유지를 하는 다른 민감속성과의 조합은 취약함
- KLT프라이버시 모델을 적용시, k-익명성의 취약한 부분을 보강하여 위험성을 현저하게 낮출 수 있음
- MAS 기법의 경우 연결형 유일성 검사에서 모든 속성 조합에 대해 위험 레코드가 0건 검출되어 연결공격에 대한 위험이 낮음
- 추가적으로 k-익명성 기법의 경우 비식별화 A-B 연계 결과에서 유일한 레코드가 비식별

화 결과 A세트에서도 유일한 레코드 값을 가졌다면, 임시를 통해 원본을 찾아낼 수 있으며, 제공받은 연계 데이터 셋을 통해 대응되는 해당 레코드의 B사 측의 속성값이 유출되는 문제가 있음

- MAS 비식별화의 경우는 N:1 마이크로집계 방식으로 대응되는 원본을 식별하기 어려우며, 연계되어 제공된 B사의 속성값 또한 원본과 동일한 값이 아닌 비식별화 처리된 속성값이므로 레코드 재식별 가능성이나 속성값 유출의 위험성이 상대적으로 낮은 것으로 판단됨

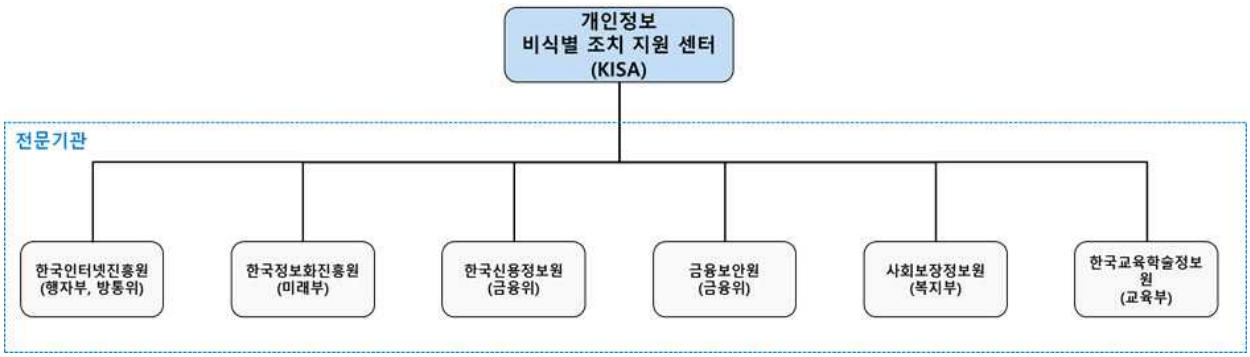
6. 비식별 조치 적정성 프로세스 실증

가. 적정성 평가

(1) 적정성 평가란

○ 비식별 조치를 통해 데이터가 다른 정보와의 결합, 다양한 추론 기법 등을 통해 개인이 식별될 우려를 발생 할 수 있으며, 이에 따라 개인정보 보호 책임자의 책임하에서 외부 전문가가 참여하는 비식별 조치 적정성 평가단(이하 평가단)을 구성, 개인의 식별 가능성에 대한 엄격한 평가를 통해 비식별 조치의 적정성의 적합 여부를 평가하는 과정임

○ 비식별 조치에 대해서는 산업분야별 관련부처에서 지정한 전문 기관을 통해 지원 받을 수 있으며 통신 산업의 전문기관인 한국인터넷진흥원(KISA)과 한국정보화진흥원(NIA)의 두 기관의 지원을 통하여 위해서 적정성 평가를 각각 수행하였음



[그림 3-29] 비식별 조치 지원 전문 기관

○ 과제를 통해 비식별 조치를 취한 13개의 데이터 중 6건의 데이터를 외부로 유통 실증을 위해 전문기관을 통해 추천 받은 외부 전문가 Pool을 포함하여 정보보호 책임자의 주관하에 평가 위원 구성하여 적정성 평가를 시행하였으며, "적합" 평가를 받은 비식별 데이터 중 H보험과의 1건의 정보 집합물 결합을 추진하였고 S신용평가로의 유통 등 개인정보 비식별 조치 가이드 라인의 전 과정에 대한 실증을 추진하였음

데이터 구분	목적	전문 기관	일정
이동전화 통화 및 수납 정보	H보험 결합 전 평가	한국인터넷진흥원	'17.2월
이동전화 통화료 수납과 보험 및 신용대출 정보	H보험 결합 후 평가	한국인터넷진흥원	'17.3월
서울시 이동전화 통화 및 수납 정보	S신용평가 유통용	한국인터넷진흥원	'17.3월
서울지역 20대 이동전화 통화 및 수납 정보	온라인 유통용(웹게시)	한국정보화진흥원	'17.4월
사회 약자 시간대별 체류지 정보	온라인 유통용(웹게시)	한국정보화진흥원	'17.4월
외국인 가입자 시간대별 체류지 정보	온라인 유통용(웹게시)	한국정보화진흥원	'17.4월

(2) 비식별 조치 결과

○ 적정성 평가를 수행할 비식별 데이터 셋은 유통 실증의 유효성 검증을 위해 '16년 12월 가입자 중 마케팅활용 선택 동의를 받은 가입자를 대상으로 이동전화 통화 및 수납 정보를 추출하여 비식별 조치 결과 데이터는 원본 대비하여 약 96%의 비식별 데이터가 생성하였으며, 온라인 유통(웹게시)용 데이터는 샘플데이터만을 추출하여 최대 36.8만명에서 6.3만명의 샘플의 비식별 조치 데이터를 생성하였음

데이터 구분	목적	데이터 크기 (건)	
		원본	비식별 조치 데이터
이동전화 통화 및 수납 정보	H보험 결합 전 사전 평가	18,019,816 (24개 항목)	18,019,816 (22개항목, 결합KEY포함)
이동전화 통화료 수납과 보험 및 신용대출 정보	H보험 결합 후 평가	2,185,596 (42개항목)	2,149,421 (42개항목)
서울시 이동전화 통화 및 수납 정보	S신용평가 유통용	3,909,347 (43개항목)	3,754,020 (68개항목)
서울지역 20대 이동전화 통화 및 수납 정보	온라인 유통용 (웹게시)	536,423 (28개 항목)	368,093 (19개 항목)
사회 약자 시간대별 체류지 정보	온라인 유통용 (웹게시)	885,433 (42개 항목)	183,610 (39개 항목)
외국인 가입자 시간대별 체류지 정보	온라인 유통용 (웹게시)	319,110 (41개 항목)	63,424 (38개 항목)

○ 비식별화 조치가 적절한 수준에서 데이터 유실을 최소화 하기 위해서는 반복적인 비식별화를 통해 다양한 비식별 조치 방법 중 최적의 방안 도출이 필요함. 해당 산업의 이해가 높으며 데이터 유실을 최소화 할 수 있도록 하는 컨설팅 지원 및 교육 등 의 전문기관의 역할이 요구됨

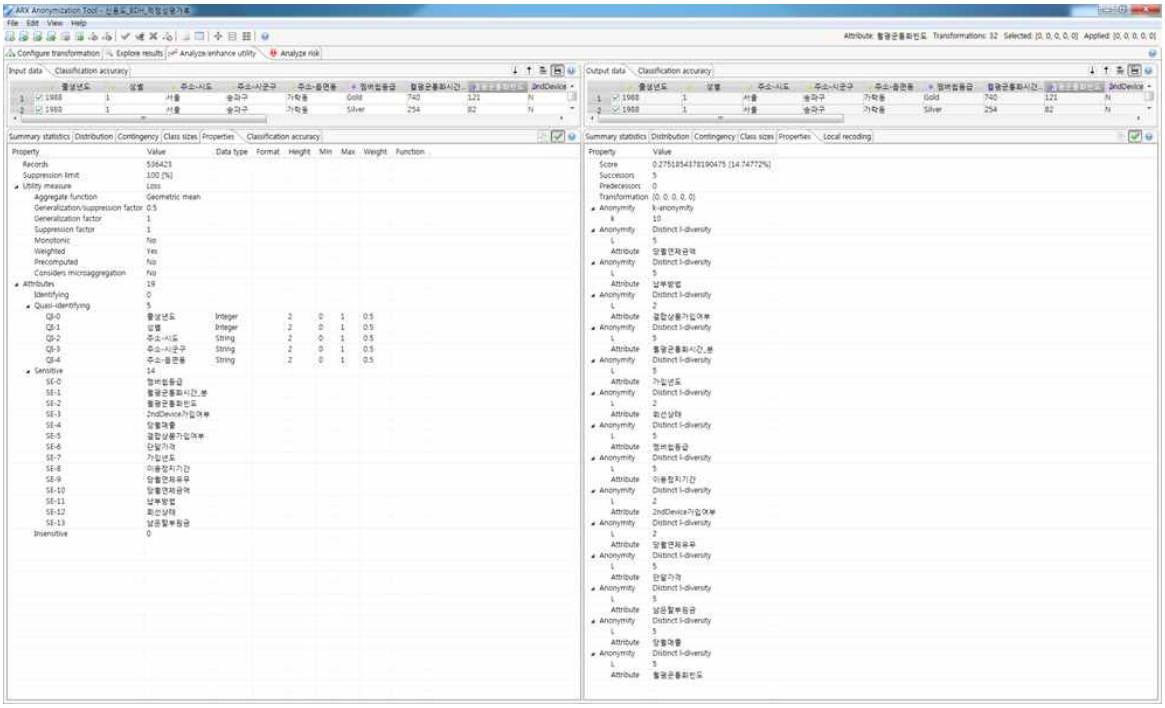
○ 외부 유통 실증 데이터의 경우 참여기관인 이지서티 k-익명성 기법와 그리즐리 MAS 솔루션을 적용하여 비식별 조치를 취하였으며, 온라인 유통용의 경우 오픈소스인 ARX 솔루션으로 비식별 조치를 취하였음

데이터 구분	목적	비식별S/W	속성자 정보 구분	비식별 조치 기준(결과)
이동전화 통화 및 수납 정보	H보험 결합 전 사전 평가	IDENTITY SHIELD	준식별자(1) 일반속성정보(21)	k-익명성(114,463) l-다양성(2)
이동전화 통화료 수납과 보험 및 신용대출 정보	H보험 결합 후 평가	IDENTITY SHIELD	준식별자(3) 일반속성정보(39)	k-익명성(5) l-다양성(2)
서울시 이동전화 통화 및 수납 정보	S신용평가사 유통용	그리즐리 MAS	준식별자(4) 일반속성정보(64)	추상화(2) k-익명성(3) l-다양성(2)
서울지역 20대 통화 및 수납 정보	온라인 유통용 (웹게시)	ARX 3.5.1	준식별자(4) 일반속성정보(15)	k-익명성(29) l-다양성(5)
사회 약자 시간대별 체류지 정보	온라인 유통용 (웹게시)	ARX 3.5.1	준식별자(7) 일반속성정보(32)	k-익명성(22) l-다양성(5)
외국인 가입자 시간대별 체류지 정보	온라인 유통용 (웹게시)	ARX 3.5.1	준식별자(6) 일반속성정보(32)	k-익명성(10) l-다양성(5)

○ 비식별 조치 과정에서 일반속성정보의 경우 L 다양성 값의 적용 외 추론이 불가하도록 백만원 이상의 금액으로 절상, 특이 금액의 경우 삭제 또는 ○○이상으로 변경 등의 범주화(범위방법)을 사용하고 범주화된 값을 Random swapping의 방법 등을 사용하여, 재식별의 위험성을 낮추었으나 가이드에 명시된 기준이 없어 적용 수준 차이가 발생할 수 있음

Summary statistics | Distribution | Contingency | Class sizes | Properties | Local recoding

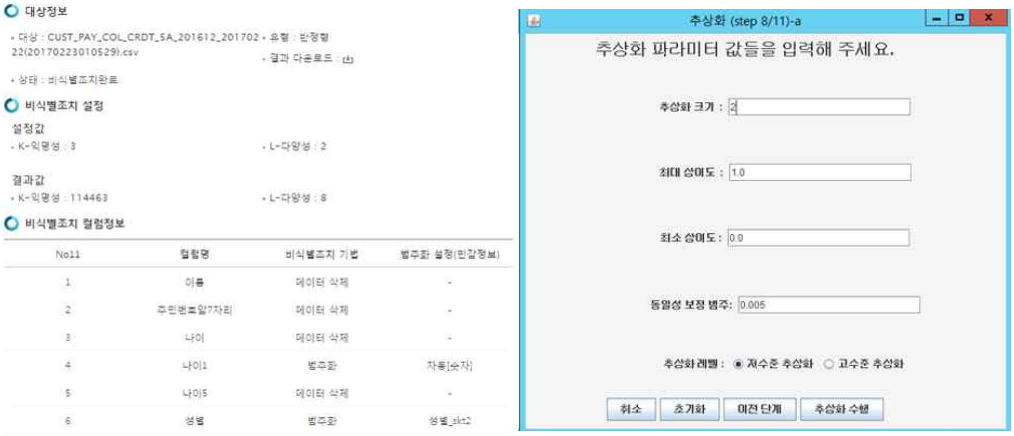
Measure	Including outliers	Excluding outliers
Average class size	340.80241 (0.06353%)	219.92816 (0.06357%)
Maximal class size	190476.0 (35.50854%)	1246 (0.36017%)
Minimal class size	29 (0.00541%)	29 (0.00838%)
Number of classes	1574	1573
Number of records	536423	345947.0 (64.49146%)
Suppressed records	190476.0 (35.50854%)	0



출생년도	성별	주소-시도	주소-시군구	주소-읍면동	가입년도	멤버쉽등급	2ndDevice가입여부	결합상품가입여부	월평균통화시간_분	월평균통화빈도	당월납부금액	당월연체유무	당월연체금액	회선상태	이용정지기간	납부방법	단말가격
1988	남	서울	송파구	가락동	2006	Gold	N	Y	[637:740]	[119:121]	[48011:49090]	Y	[1:100000]	사용중	1년	입금전용계좌	[0:500000[
1988	남	서울	송파구	가락동	2004	Silver	N	N	[248:254]	[82:82]	[41821:42860]	N	0	사용중	1개월미만	지로납부	[600000:700000[
1988	남	서울	송파구	가락동	1999	Gold	N	Y	[637:740]	[122:125]	[24491:26010]	N	0	사용중	2개월	은행자동납부	[600000:700000[
1991	여	서울	송파구	거여동	2012	Gold	N	Y	[105:107]	[0:63]	[48011:49090]	N	0	사용중	1개월	은행자동납부	[0:500000[
1991	여	서울	송파구	거여동	2012	Gold	N	Y	[39:41]	[0:63]	[37381:38300]	N	0	사용중	1개월미만	카드자동납부	[0:500000[
1991	여	서울	송파구	거여동	2012	Gold	N	N	[320:331]	[0:63]	[53661:54870]	N	0	사용중	1개월미만	은행자동납부	[900000:1000000[
1992	여	서울	양천구	목동	2014	Gold	N	N	[410:431]	[98:99]	[49091:50190]	N	0	사용중	1개월미만	은행자동납부	[0:500000[
1992	여	서울	양천구	목동	2014	Silver	N	N	[66:68]	[0:63]	[42861:43920]	Y	[1:100000]	사용중	1개월미만	은행자동납부	[0:500000[
1992	여	서울	양천구	목동	2014	Silver	N	N	[11:14]	[0:63]	[50191:51380]	N	0	사용중	1개월미만	은행자동납부	[700000:800000[
1994	남	서울	구로구	구로동	2011	없음	N	Y	[345:358]	[164:168]	[28561:29890]	Y	[1:100000]	사용중	1개월미만	입금전용계좌	[500000:600000[
1994	남	서울	구로구	구로동	2011	없음	N	Y	[63:65]	[0:63]	[10741:12900]	N	0	정지	1년	은행자동납부	[0:500000[
1994	남	서울	구로구	구로동	2011	없음	N	N	[140:143]	[116:118]	[70631:71740]	N	0	사용중	1년	은행자동납부	[0:500000[
1994	남	서울	강서구	화곡동	2015	없음	N	N	[0:0]	[0:63]	[59371:60510]	N	0	사용중	1개월미만	은행자동납부	[0:500000[
1994	남	서울	강서구	화곡동	2015	일반	N	Y	[30:32]	[0:63]	[28561:29890]	N	0	사용중	1개월미만	은행자동납부	[900000:1000000[
1994	남	서울	강서구	화곡동	2015	일반	N	Y	[457:486]	[147:153]	[103441:105410]	Y	[1:100000]	사용중	1개월미만	입금전용계좌	[0:500000[
1995	남	서울	강북구	미아동	2016	일반	N	N	[3:6]	[0:63]	[29891:31070]	Y	[1:100000]	사용중	1개월미만	은행자동납부	[0:500000[
1995	남	서울	강북구	미아동	2016	없음	N	N	[11:14]	[0:63]	[46951:48010]	N	0	사용중	1개월미만	은행자동납부	>1000000
1995	남	서울	강북구	미아동	2016	일반	N	N	[487:523]	[0:63]	[37381:38300]	N	0	사용중	1개월미만	은행자동납부	[0:500000[
1995	남	서울	강북구	미아동	2016	일반	N	N	[0:0]	[83:83]	[0:0]	N	0	사용중	1개월미만	NA	[0:500000[
1995	남	서울	강북구	미아동	2016	없음	N	N	[0:0]	[0:63]	[0:0]	N	0	사용중	1개월미만	NA	>1000000
1995	남	서울	강북구	미아동	2016	없음	N	N	[0:0]	[0:63]	[0:0]	N	0	사용중	1개월미만	NA	[0:500000[
1996	남	서울	강동구	암사동	2014	Silver	N	N	[57:59]	[0:63]	[1:7580]	Y	[1:100000]	정지	3개월	은행자동납부	[0:500000[
1996	남	서울	강동구	암사동	2014	일반	N	N	[75:77]	[0:63]	[26011:27280]	N	0	사용중	1개월	은행자동납부	[0:500000[
1996	남	서울	강동구	암사동	2014	없음	N	Y	[3:6]	[0:63]	[29891:31070]	N	0	사용중	1개월미만	카드자동납부	[800000:900000[
1997	남	서울	강남구	대치동	2013	일반	N	Y	[0:0]	[0:63]	[14421:16460]	N	0	사용중	6-12월	카드자동납부	[900000:1000000[
1997	남	서울	강남구	대치동	2013	Silver	N	N	[21:23]	[0:63]	[39961:40860]	N	0	사용중	1개월미만	은행자동납부	[0:500000[
1997	남	서울	강남구	대치동	2013	Silver	N	Y	[140:143]	[78:78]	[37381:38300]	N	0	사용중	1개월미만	카드자동납부	[0:500000[
1997	여	서울	송파구	잠실동	2016	없음	Y	N	[173:177]	[0:63]	[46951:48010]	N	0	사용중	1개월미만	카드자동납부	[0:500000[
1997	여	서울	송파구	잠실동	2016	없음	N	Y	[39:41]	[0:63]	[27281:28560]	N	0	사용중	1개월미만	은행자동납부	[900000:1000000[
1997	여	서울	송파구	잠실동	2016	없음	N	N	[0:0]	[0:63]	[1:7580]	Y	[200001:300000]	정지	1개월미만	입금전용계좌	[0:500000[
1997	여	서울	강서구	화곡동	2016	없음	Y	N	[0:0]	[0:63]	[14421:16460]	Y	[1:100000]	사용중	1개월미만	지로납부	[0:500000[
1997	여	서울	강서구	화곡동	2016	없음	Y	N	[0:0]	[0:63]	[7581:10740]	Y	[1:100000]	사용중	1개월미만	지로납부	[0:500000[
1997	여	서울	강서구	화곡동	2016	없음	N	Y	[93:95]	[0:63]	[35281:36430]	Y	[1:100000]	사용중	1개월미만	은행자동납부	[500000:600000[

〉1000000 :1000,000 초과, [0:500000[: 0이상 ~ 500,000 미만, [200001:300000] : 200,001 이상 ~ 300,000이하, 〈20 : 20 미만, 〉=80 : 80이상

[그림 3-30] 수납 정보 비식별 조치 결과(ARX)



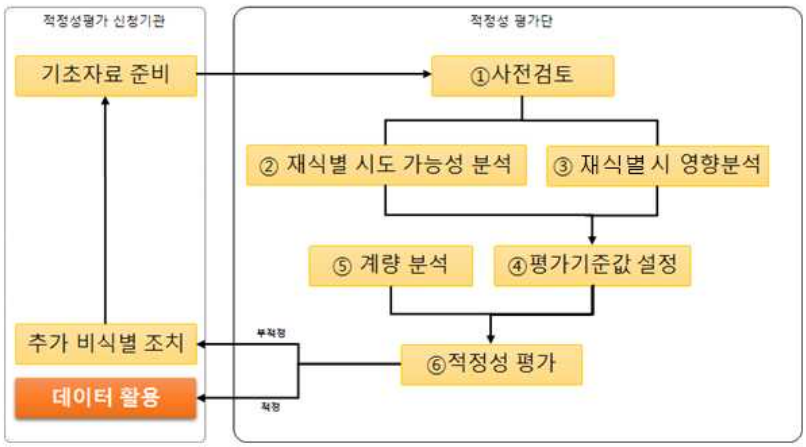
[그림 3-31] 수납 정보 비식별 조치 결과(IDENTITY-SHIELD, MAS)

(3) 적정성 평가 수행 결과

○ 적정성 평가는 적정성 평가 신청기관에서 기초자료의 준비 후 사전검토 단계부터 적정성 평가까지의 6단계를 거치게 되며, ②재식별 시도 가능성 분석, 평가 대상 데이터의 ③재식별 시 영향분석 결과를 바탕으로 ④평가 기준 값을 설정하여 적정성 평가 기관에서 취한 비식별 조치 수준과의 비교를 통해 "적합"여부를 최종적으로 판단하게 되어있음

○ ③재 식별 시 영향분석의 경우, 데이터의 유출 이후 재식별이 되었다는 경우를 가정하여 평가를 하게되면, 이는 비식별조치 후 적정성 평가를 한 신청 기관이 개인정보보호 수준이 높고, 재식별 시도 가능성이 거의 없는 경우에도 침해 위험이 낮은 수준의 평가가 불가하며 변별력있는 평가가 되지 않게 나타남. 이에 따라 과제 중 적정성 평가 시 개인정보 비식별조치 지원센터인 한국인터넷진흥원이 함께 참여하여 "재식별 시 영향분석"을 "유출 시 영향분석"으로 변경 개선하도록 방안을 수립하였으며 하반기 중에 평가과정에서 나타는 불필요한 부분을 삭제한 간소화된 양식으로 변경될 수 있도록 하였음.

<평가 세부 절차>



○ 총 6번의 데이터를 적정성 평가 신청을 하여 평가를 받아본 결과, 동일한 내용을 중복적으로 준비 및 평가를 해야 하는 경우가 발생하고 있으며 주기적, 반복적인 비식별 데이터를 생성하고 분석하는 경우 최초 평가 이후 추가 평가 시 평가 절차 간소화가 필요함

(4) 데이터 결합 수행 결과

○ 비식별 자료의 결합을 위해서는 A社와 B社가 같은 알고리즘을 적용하여 “주민등록번호”가 아닌 식별자를 임시 대체키로 전환하고, 결합대상 정보집합물도 비식별 조치 및 적정성 평가 수행함

[그림 3-32] 정보집합물 결합 절차

○ 실제 비식별 정보의 결합 후 분석활용까지는 최소 3개월 이상이 소요되었으며, 경험 부재의 신규 기관의 경우 각 세부 프로세스 진행 단계에서 오류 및 추가 비식별 조치가 발생할 경우 추가 시간 소요가 발생할 수 있음. 이를를 고려하여 데이터 제공 표준 협약 양식 배포 및 적정성 평가 기준 안내와 같은 지원을 통해 평가 기간 단축이 필요함

단계	세부 절차	진행 기간
결합 기관간 사전 협의	기관 간 보유 데이터 현황 파악 결합 데이터 필드 정의 결합 임시 대체 KEY 및 생성 알고리즘 정의 비식별 자료 제공 협약서 체결	‘17.1월
결합 전 적정성 평가	비식별 조치 및 데이터 생성 적정성 평가 신청 평가 및 결과 수령	‘17.2월
사전 결합	사전 결합 신청 결합 임시 대체 KEY 반출 및 결합 사전 결합 결과 수령	‘17.3월
데이터 결합	결합 신청 결합 데이터 반출 및 결합 결합 결과물 수령	‘17.3월
결합 후 적정성 평가	비식별 조치 및 데이터 생성 적정성 평가 신청 평가 및 결과 수령	‘17.4월
결합 완료 및 데이터 분석	분석 및 활용 (분석 과정에서 데이터 활용에 필요이상의 데이터 정보손실 발생 시 과정 반복 불가피)	‘17.4월

○ 금융과 통신 정보의 집합물 결합 시 비식별 조치 및 결합과정에서의 데이터 손실이 불가피하나, 금융권 등 타 결합 사례의 경우 결합 및 비식별 조치 이후 전체 데이터 모수의 결합률 10~20% 수준으로 효용성 저하의 우려가 있으며, 가입자가 많은 이번 이동통신사의 데이터의 경우 비식별 조치 및 결합과정에서의 47.6%의 결합률을 나타나 통신 외 산업과의 결합 후 활용도가 높게 나타남

구분		데이터	값
결합 전	결합 대상 정보집합물 수량	SKT 통신료 수납정보	1,802만(건)
		H보험 보험 및 신용대출 정보	459만(건)
결합 후	결합 DB 수량	SKT 및 H보험 결합 집합물	2,185,596(건)
	결합율(잔존율)	SKT 통신료 수납정보	12.1%
		H보험 보험 및 신용대출 정보	47.6%

데이터 산출 기준 : 2016년 12월 정상 가입자

○ 임시 대체키 생성 시 통신사업자의 식별자인 이동전화번화의 경우 보험사에서 현행화가 되어 있지 않고 있어 결합률이 저하될 수 있는 것으로 파악되었며, 결합 기관들이 동시에 활용 가능한 유일 식별자를 보유하고 있지 않아 생년월일, 성별, 이름의 준식별자 조합을 통해 식별자를 생성하였음, 이 경우 동명 이인의 중복고객 발생 가능성이 있으며, 오류 가능성을 낮추기 위해 삭제시 결합률이 낮아지는 영향을 발생하기 때문에, 연계정보(CI, Connecting Information)와 같이 고유 값의 일부를 조합하는 등 별도의 중복 가능성을 낮추고 결합률을 높일 수 있는 임시 대체키 생성에 대한 가이드가 필요함. 본 과제 중에는 CI 활용에 대한 법무적 검토 및 데이터 추가 추출시 소요되는 기간을 고려하여 제외하여 추진하였으며, 사전 가이드 등이 있을 경우 다양한 임시 대체키를 생성하여 사전 결합율을 확인 할 수 있다면 활용 가능성을 높일 수 있음

○ 이종의 산업 간의 정보집합물 결합 시 양사간에 동일하게 보유 중이며 활용할 수 있는 식별자가 제한적인 측면이 있어, 데이터의 활용도를 높일 수 있도록 임시 대체키 생성 방안 및 “임시 대체키”의 생성 및 사전 결합률에 대해서는 현재의 가이드에 생성 및 신청 절차를 추가할 필요가 있음

○ 결합 과정 중 결합 절차에 신청서 및 확인서에 회사 직인, 결합 신청을 위한 방문 접수 및 방문 수령 등의 활성화를 위한 불편한 점이 발견되었으며, 다수의 신청서 및 확인서 통합, 기관간의 데이터 송수신 방안에 대한 개선 필요함

7. 법 제도 및 규제개선 사항

가. 배경

2016년 6월 개인정보 비식별조치 가이드라인이 발표된 후에도 개인정보처리사업자들의 비식별정보 생성 및 유통이 활발하게 이루어지고 있지는 못함

그 원인은 비식별기술의 미발달, 법 제도의 미비점, 사회의 부정적 인식의 유지 등에서 찾아볼 수도 있을 것인데, 그 중에서도 현행 법 제도의 미비점이 큰 비중을 차지할 것으로 보임

금번 과제 수행과정에서 과제수행의 대상기준 선정 및 비식별 목적물인 데이터 선정기준, 비식별 과정 및 적절성평가 과정 등에 있어 법제도 상의 불명확한 부분들이 발견되었는데, 과제 수행 결과의 안전성을 위하여 관련 해석상 모호한 부분은 가장 보수적인 해석을 기준으로 하여 과제를 진행

이하에서는 비식별정보 유통 실증을 위해 현행 법 제도의 미비점들을 살펴보고, 개선 필요사항을 정리함

나. 비식별정보 생성의 대상 데이터 및 범위

개인정보보호법 및 정통망법 기타 비식별조치 가이드라인에서는 비식별정보 생성의 대상이 되는 개인정보의 범위에 대하여 규정하고 있지 않으며, '비식별조치'하는 과정이 개인정보의 '이용'에 해당하는 것인지 여부에 대하여도 논하고 있지 않음

원칙적으로 사업자는 수집과정에서의 불법이 있는 경우가 아니라면 자신이 수집한 개인정보를 비식별조치 대상으로 할 수 있으며, 그 범위에 제한은 없음

비식별조치 자체는 개인정보 "이용"이 아니므로 비식별조치를 수행하는 것에 대한 동의를 득할 필요는 없음*

* 비록 비식별가이드라인의 발표로 폐기된 안내서이지만, 미래부, 한국정보화 진흥원, 빅데이터전략센터에서 발간한 개인정보 비식별화 기술활용안내서(2015.6.10) P.23에서는 구체적인 논거는 기재하지 않았으나, "비식별화 처리에 대한 정보 주체의 동의가 필요하다고 볼 경우 빅데이터 사업의 육성 및 효과적인 데이터 분석을 저해하는 면이 있을 뿐아니라, 법적으로도 비식별화 자체에 대해 정보주체의 동의를 요한다고 볼 수 없음"이라고 명확히 의견을 제시하고 있음

이 문제 또한 비식별정보의 유통에 대하여 사회적 공감대가 완벽하게 형성되어 있지 않은 상황에서 비식별조치를 할 수 있는 원천 대상정보의 범위에 대한 논란을 없애기 위해서 비식별정보의 정의를 법정화하여 명확히 할 필요가 있음(아. 개선안 참조)

다. 비식별정보의 개인정보성

현행 가이드라인에서는 "이 가이드라인에 따라 정보주체를 알아볼 수 없도록 비식별조치를 적정하게 한 비식별정보는 개인정보가 아닌 것으로 추정"된다고 규정하고 있음

정보주체를 알아볼 수 없도록 조치한 정보라면 이미 개인정보가 아니어야 할 것임. 그럼에도 가이드라인에서는 개인정보가 아닌 것으로 '추정'한다고 규정하여 사업자들의 비식별정보 유통에 대한 의지를 매우 약하게 만들고 있음

더욱이 법적인 효력이 있지 않은 가이드라인에서 '추정'을 인정한다 하더라도 이는 법률상 추정이 될 수 없다는 한계가 있음. 더욱이 개인정보 여부에 대한 입증은 사실상 개인정보처리자가 하게 될 것이어서 이러한 추정의 의미가 크지는 않음

다만, 행정처분에 대한 신뢰의 원칙 등에 기초하여 가이드라인에서 정한 프로세스를 따랐을 경우 '추정'은 행정처분에서의 책임면제는 가능할 것이고, 또한 수사기관 등에서의 고려가 충분히 있을 것이라는 정도의 범위에서 의미가 있는 것으로 보임

이 문제 또한 비식별정보의 유통에 대하여 사회적 공감대*가 완벽하게 형성되어 있지 않은 상황에서 비식별조치에 대한 논란을 없애기 위해서 비식별정보의 정의를 법정화하여 명확히 할 필요가 있음(아. 개선안 참조)

* 비식별가이드라인 발표 이후에도 시민단체나 NGO들에서는 비식별정보 자체에 대한 법정화를 반대한다는 의견을 지속적으로 내고 있는 상황으로, 법정화되기 전까지는 동일한 논란이 지속될 수 밖에 없을 것으로 보임 http://news.inews24.com/php/news_view.php?g_serial=993063&g_menu=020200&rrf=nv

라. 비식별 가이드라인에 따른 적정성 평가의 의미

현행 가이드라인에서는 가이드라인에서 정한 절차에 따르는 비식별정보에 대하여서만 개인정보가 아닌 것으로 추정한다고 규정하고 있음

그런데, 가이드라인에서는 비식별의 기준으로 k-익명성 모델만을 기초 기준으로 하고 있을 뿐, 적정성 평가의 구체적인 기준에 대하여는 사실상 사업자에게 모두 맡기고 있는 실정*으로, 결국 비식별가이드라인이 존재한다고 하더라도 비식별정보의 유통에 대한 risk와 책임이 사업자에게 귀속될 수 밖에 없는 상황임

* 현재 지정된 산업분야별 전문기관(한국인터넷진흥원, 한국신용정보원, 금융보안원, 사회보장정보원, 한국정보화진흥원)에서는 적정성평가에 대한 세부지표를 작성하여 사업자의 적정성평가시 사용할 수 있도록 지원하는 등의 노력을 하고는 있으나, 확정된 세부지표가 완성이 되지 않았고 각 전문기관 별로 표준화 되어있지도 않아 활용도가 떨어질 수 밖에 없음

비식별정보화 기술은 지속 발전해나갈 것이고, 실제 금번 과제 수행과정에서도 새로운 형태의 비식별기술(MAS 비식별 기법)을 적용해본 바 있으나, 이렇게 새로운 기법에 따라 비식별조치한 경우의 적정성평가 방식이 k-익명성 이라는 일률적인 기준으로 판단하는 것이 적절한 것인지에 대한 의문이 존재함

이러한 상황에 대해 가인드라인은 개인정보를 비식별정보화하는 하나의 방안을 제시한 것일 뿐이고, 따라서, 사업자가 자신의 방식으로 비식별정보화 하고 가이드라인에 따른 적정성 평가를 거치지 않았다 하더라도 추후 사업자가 스스로 비식별정보가 아님을 충분히 입증할 수 있다면 유통하여도 된다고 해석할 수도 있으나, 이는 결국 가이드라인의 존재의미가 몰각되는 결론에 이르게 되므로 적절하지 않음

따라서, 이러한 문제는 법에서 "비식별정보"에 대하여 규정하면서 해당 "비식별조치"의 적정성평가의 기준을 대통령령 등에 위임하고 대통령령에서는 기본기준을 규정하되 세부내용을 다시 하위규정으로 재위임을 하여 법적 근거를 마련하되 세부기준마련에 유연성을 가질 수 있도록 하는 해결이 필요할 것임(아. 개선안 참조)

마. 가이드라인에 따른 비식별정보와 개인정보보호법상의 “통계”정보

가이드라인에 따른 비식별정보와 개인정보보호법상의 동의없이 제3자제공이 가능한 “통계작성 및 학술연구 등의 목적을 위하여 필요한 경우로서 특정 개인을 알아볼 수 없는 형태로 개인정보를 제공하는 경우”의 정보(개인정보보호법 제18조 제2항 제4호)는 비식별조치를 한다는 점에서는 동일하나 절차적으로 볼 때 “통계”정보의 경우 적정성평가를 거치지 않는다는 차이가 있음

두 개념의 구분이 큰 의미가 없다는 의견*이 있으나, 개인정보를 동의없이 이용하거나 제3자 제공하는 경우 개인정보보호법 및 정통망법에서 엄중한 형사처벌을 하고 있다는 점을 고려할 때 비식별조치를 한 정보의 이용에 대하여 법에서 명확히 정보처리사업자를 면책하는 내용으로의 입법이 필요함

* 두 정보의 관계에 대하여는 i) 개인정보가 아닌 것으로 추정되는 점에서 비식별정보와 동일하지만 이용목적에 있어 통계작성 및 학술연구 등을 위해 국한된다는 차이가 있다는 의견, ii) 개인정보보호법은 여전히 개인정보에 해당한다고 규정하고 있어 비식별가이드라인에 따른 비식별정보와는 구분된다는 의견, iii)통계정보는 정보의 이용목적에 관한 표현이고, 비식별정보는 그 정보의 성질에 관한 것으로 접근 방식이 다른 것이고 때로는 동일한 개념이고 때로는 다른 개념일 수 있으며 동의없는 제3자 제공이 가능한 것인지에 대한 범위에서 큰 차이가 없고 논의의 실익이 크지 않다는 의견이 있음

〈현행법상 벌칙규정〉

개인정보보호법	정보통신망법
제71조(벌칙) 다음 각 호의 어느 하나에 해당하는 자는 **5년 이하의 징역 또는 5천만원 이하의 벌금**에 처한다. 1. (생략) 2. **제18조제1항·제2항**, 제19조, 제26조제5항 또는 제27조제3항을 위반하여 **개인정보를 이용하거나 제3자에게 제공한 자 및 그 사정을 알면서도 영리 또는 부정한 목적으로 개인정보를 제공받은 자**	제71조(벌칙) ①다음 각 호의 어느 하나에 해당하는 자는 **5년 이하의 징역 또는 5천만원 이하의 벌금**에 처한다. 1. 2. (생략) 3. 제24조, **제24조의2제1항 및 제2항** 또는 제26조제3항(제67조에 따라 준용되는 경우를 포함한다)을 **위반하여 개인정보를 이용하거나 제3자에게 제공한 자 및 그 사정을 알면서도 영리 또는 부정한 목적으로 개인정보를 제공받은 자**

바. 비식별정보 “결합”에 있어서의 임시대체키 문제

가이드라인의 비식별조치는 식별자를 충분하게 제거하여 개인식별이 불가능하도록 k-익명성 기준을 충족하여야만 적정성평가를 통과할 수 있도록 하고 있음

그런데, 서로 다른 사업자가 보유하고 있는 정보집합물을 결합하고자 하는 경우, 그 결합을 위하여 각 정보별로 매칭키를 부여해야하고 이 경우 결국 현행법상 매칭키는 그 자체로는 개인정보가 아니지만 “쉽게 결합하여 특정개인을 알아볼 수 있는 정보”에 해당될 소지*가 매우 높음

* 실제로 가이드라인에서도 그 위험성에 대해서 설명하면서, 전문기관을 통하여서 “결합”을 진행하도록 할 필요가 있음을 안내

○ 빅데이터 분석에 활용하기 위해 서로 다른 사업자가 보유하고 있는 정보집합물을 결합하는 경우 개인별로 부여된 식별자가 매칭키로 사용됨
- 이 경우, 정보주체를 알아볼 수 있는 식별자 그 자체를 매칭키로 사용하는 것은 현행법 위반소지

○ 따라서, 정보집합물 간 결합·분석을 위해서는 결합과정에서만 임시로 매칭키 역할을 하는 '임시대체키'의 활용이 필요

○ 임시대체키를 활용한 결합을 허용하는 경우에는 무분별한 결합을 통한 개인정보 침해 소지를 방지하기 위해 전문기관에서만 결합하도록 하는 등 지원 및 관리체계 필요

현재 가이드라인에서는 매칭키로 "임시대체키"를 발급하여 사용하도록 하면서, 결합은 전문기관을 통하여 하도록 함. 이를 그림으로 설명하면 아래와 같음

그러나, "임시대체키"는 기본적으로 개인정보를 매칭할 수 있는 "매칭키"이고, 단지 이를 임시로 새로이 발급했다는 점 외에는 여전히 '쉽게 결합하여 특정개인을 알아 볼 수 있는 정보'에 해당할 수 있게 되고, 대체키에 묶여있는 정보 전체를 "비식별정보"라고 보기 어려울 수 있음

개인정보성이 완전하게 탈락되지 않은 정보라면 이 상태로 전문기관에게 전달하는 것이 동의 없이 제3자에게 개인정보를 전달하는 것이 될 위험이 있음

이에 대하여 "전문기관"이 안전하고 적정히 관리한다는 전제라면 비식별정보로 봐야하며, 가이드라인에 있는 이상 행정처분은 신뢰보호 원칙 등에 근거하여 면책이 될 수 있을 것이고, 법원 등에서도 가이드라인에서 정한 절차를 따른 경우 비식별정보로 인정될 여지가 충분히 있다는 견해도 있으나, 비식별정보의 유통 활성화를 위하여는 해당 내용은 법에서 명시적으로 개인정보에서 제외하여 줄 필요가 있을 것임

즉, 이러한 문제 또한 법에서 명시적으로 "비식별정보"의 개념을 도입하지 않았기 때문에 발생되는 것이고, 법상 비식별조치의 개념에 '임시대체키'에 관한 근거를 마련하여 입법적으로 해결이 되어야 함(아. 개선안 참조)

※ 2015년 개정된 일본의 개인정보보호법에서는 "익명가공정보"의 정의를 신설하였는데, 그 내용 중에는 우리가 참고할만한 난수 개념의 '대체키'에 대한 내용이 포함되어 있음

일본 개인정보보호법 제2조 제9항
이 법률에서 “익명가공정보”란 다음 각호에 해당하는 개인정보 구분에 따라서 해당 각호에 규정된 조치를 취하여 특정개인을 식별할 수 없도록 개인정보를 가공하여 얻는 개인에 관한 정보이며, 해당 개인정보를 복원할 수 없도록 한 것을 말한다.

1. 성명, 생년월일, 그 밖의 기술(記述) 등에 의해 특정 개인을 식별할 수 있는 것(다른 정보와 용이하게 대조·확인하여 특정 개인을 식별할 수 있도록 한 것도 포함)은 그 일부를 삭제(해당 일부를 복원할 수 있는 규칙성을 가지지 않는 다른 내용으로 대체하는 것을 포함)
2. 개인식별부호가 포함된 것은 개인식별부호를 전부 삭제(해당 일부를 복원할 수 있는 규칙성을 가지지 않는 다른 내용으로 대체하는 것을 포함)

사. 비식별화 소프트웨어정보 “결합”에 있어서의 임시대체키 문제

현재 시중에 상용화된 비식별화 S/W의 종류는 아래와 같으며, 비식별화 S/W는 비식별정보 유통 활성화에 맞춰 지속 개발될 것으로 예상됨. 이 중에는 공개 소프트웨어도 있고, 상업용 S/W로 출시된 것이 있음

S/W	공개여부
ARX Data Anonymization Tool UDT Anonymization Toolbox Cornell Anonymization Toolkit TIAMAT SECRETA Open Anonymizer ANON tool uArgus sdcMicro	공개 S/W
PARAT analytic DID IDENTITY SHIELD	상업용

사업자가 비식별화 S/W를 직접 보유하고 직접 조치를 하는 경우에는 비식별조치 자체가 이용에 해당하는 것이 아니라면, 그 외 특별한 법적인 문제는 발생하지 않을 것인데, 모든 사업자가 비식별화 S/W를 직접 보유하고 이용할 수 있는 것은 아니고 비식별기술 적용을 제3자에게 위탁하여 처리하는 경우 이에 대해 별도의 동의가 필요한 것이 아닌지에 대한 의문이 발생함

현행 개인정보보호법에서는 개인정보의 처리 위탁에 관하여 별도의 동의를 받지 않아도 가능하도록 하고 있는 반면, 정통망법에서는 원칙적으로 별도의 동의를 받아야 하고 그것이 계약을 이행하고 이용자의 편익을 증진하고자 하는 경우에만 예외적으로 동의없이 가능하도록 규정하고 있음(개인정보보호법 제26조, 정통망법 제25조)

비식별작업 S/W를 통해 개인정보를 “비식별조치”를 하는 것이 계약을 이행하고 이용자의 편의 증진 등을 위하여 필요한 경우라 해석하기는 어려움이 있고, 정통망법 규정에 따를 때에는 원칙적으로 동의*를 받아서 위탁하여야 하는 사항으로 보임

* 이에 대하여, 문언해석만을 하는 때에는 극단적으로 정보통신서비스제공자가 소송을 수행하기 위하여 법무법인에 이용자 개인정보를 위수탁하는 경우에도 일일이 정보주체로부터 동의를 받아야한다 것과 동일한 매우 불합리한 결과에 이르게 됨. 따라서 소송수행이나 비식별작업 S/W의 이용 등 합리적인 위수탁 현상에 대해서는 위 예외를 넓게 해석할 현실적인 필요성이 있다는 견해도 있음

개인정보보호법	정통망법
제26조(업무위탁에 따른 개인정보의 처리 제한) ① 개인정보처리자가 제3자에게 개인정보의 처리 업무를 위탁하는 경우에는 다음 각 호의 내용이 포함된 문서에 의하여야 한다. 1. 위탁업무 수행 목적 외 개인정보의 처리 금지에 관한 사항 2. 개인정보의 기술적·관리적 보호조치에 관한 사항 3. 그 밖에 개인정보의 안전한 관리를 위하여 대통령령으로 정한 사항 ② 제1항에 따라 개인정보의 처리 업무를 위탁하는 개인정보처리자(이하 "위탁자"라 한다)는 위탁하는 업무의 내용과 개인정보 처리 업무를 위탁받아 처리하는 자(이하 "수탁자"라 한다)를 정보주체가 언제든지 쉽게 확인할 수 있도록 대통령령으로 정하는 방법에 따라 공개하여야 한다. ③(생략) ④ **위탁자는 업무 위탁으로 인하여 정보주체의 개인정보가 분실·도난·유출·위조·변조 또는 훼손되지 아니하도록 수탁자를 교육하고, 처리 현황 점검 등 대통령령으로 정하는 바에 따라 수탁자가 개인정보를 안전하게 처리하는지를 감독**하여야 한다. ⑤ ~⑦ (생략)	제25조(개인정보의 처리위탁) ① 정보통신서비스 제공자와 .....(중략)...(이하 "개인정보 처리위탁"이라 한다)하는 경우에는 **다음 각 호의 사항 모두를 이용자에게 알리고 동의**를 받아야 한다. 다음 각 호의 어느 하나의 사항이 변경되는 경우에도 또한 같다. 1. 개인정보 처리위탁을 받는 자(이하 "수탁자"라 한다) 2. 개인정보 처리위탁을 하는 업무의 내용 ② 정보통신서비스 제공자등은 정보통신서비스의 제공에 관한 계약을 이행하고 이용자 편의 증진 등을 위하여 필요한 경우로서 제1항 각 호의 사항 모두를 제27조의2제1항에 따라 공개하거나 전자우편 등 대통령령으로 정하는 방법에 따라 이용자에게 알린 경우에는 개인정보 처리위탁에 따른 제1항의 고지절차와 동의절차를 거치지 아니할 수 있다. 제1항 각 호의 어느 하나의 사항이 변경되는 경우에도 또한 같다. ③ (생략) ④ **정보통신서비스 제공자등은** 수탁자가 이 장의 규정을 위반하지 아니하도록 **관리·감독 및 교육하여야 한다.** ⑤ ~⑦ (생략)

현행법 하에서 활용할 수 있는 방안은 결국 S/W의 라이선스를 취득하여 사업자가 직접 비식별조치를 한다거나, 혹은 비식별조치 S/W 사업자로부터 해당 업무를 수행할 수 있는 직원을 파견받아서 개인정보처리자(위탁자)의 직접적인 업무지시 및 관리감독 하에서 업무 수행을 하도록 하여 개인정보가 제3자에게 전달되지 않고 마치 직접 비식별조치를 한 것과 동일하게 처리하는 등의 방안을 활용할 수 밖에 없음

개인정보처리위탁의 경우 위탁자로서의 감독의무를 법에서 명시(및 수탁자의 행위로 인한 책임을 함께부담하도록 규정)하고 있는 이상 처리위탁에 대한 동의를 요구하는 것은 과도하며, 개인정보보호법과의 균형에도 맞지 않는다는 이유로 정부 발의로 해당 내용을 삭제하는 방향으로 정통망법 개정작업을 추진 중임

비식별조치를 모든 개인정보처리사업자가 직접 수행할 수는 없는 상황이라는 점을 고려할 때 비식별정보 유통 활성화를 위하여 조속한 입법적 조치가 필요한 부분이라고 판단됨

아. 개선안

(1) 비식별정보를 법적개념으로 승격시킬 필요 있음. 이를 위하여 정보통신망 이용촉진 및 정보보호에 관한 법률의 정의규정에 “비식별정보”를 신설필요

(2) 개인정보의 해당여부에 대하여도 정보처리자의 기준에서 접근가능성 및 결합가능성이 있는 정보만을 개인정보의 범위로 한정하여 “개인정보”정의 및 범위의 모호성을 탈피할 필요

현행	개선안
제2조(정의) 6. "개인정보"란 생존하는 개인에 관한 정보로서 성명·주민등록번호 등에 의하여 특정한 개인을 알아볼 수 있는 부호·문자·음성·음향 및 영상 등의 정보(해당 정보만으로는 특정 개인을 알아볼 수 없어도 다른 정보와 쉽게 결합하여 알아볼 수 있는 경우에는 그 정보를 포함한다)를 말한다.	**제2조(정의)** <개정>6. "개인정보"란 생존하는 개인에 관한 정보로서 성명·주민등록번호 등에 의하여 특정한 개인을 알아볼 수 있는 부호·문자·음성·음향 및 영상 등의 정보(해당 정보만으로는 특정 개인을 알아볼 수 없어도 정보처리자가 합리적으로 접근가능한 방법을 통해 다른 정보와 쉽게 결합하여 알아볼 수 있는 경우에는 그 정보처리자에 대하여는 그 정보를 포함한다)를 말한다. <신설>7. “비식별정보”란 개인정보에 포함된 개인식별 부호를 전부 삭제하거나, 해당 개인식별부호를 복원할 수 있는 규칙성을 가지지 않는 다른 내용으로 대체하여 특정 개인을 식별할 수 없도록 개인정보를 가공하여 얻은 개인에 관한 정보를 말한다.
제24조의2(개인정보의 제공 동의 등)	<신설>제24조의3(비식별처리) 정보통신서비스 제공자 등은 개인정보를 비식별처리할 때 특정 개인을 식별하는 것 및 작성에 이용하는 개인정보를 복원할 수 없도록 하기 위해 대통령령으로 정하는 기준에 따라 해당 개인정보를 비식별처리 하여야 한다.

7. 법 제도 및 규제개선 사항(연세대)

가. 규제 개선 대상 제도

(1) 개인정보 비식별 조치 가이드라인* - 비식별 조치 기준 및 지원 : 관리체계

- 2016년 국무조정실, 행정자치부, 방송통신위원회, 금융위원회, 미래창조과학부, 보건복지부가 공동으로 만든 빅데이터의 개인정보 비식별 조치 방법

나. 규제 개선안

개인정보 비식별 조치 가이드라인의 내용중 아래 2가지 개선안 제시

(1) 개선안 1 : 속성자기반 재식별 시도 방어를 위한 m-유일성 모델의 추가 활용 제안

(2) 개선안 2 :안전한 비식별 데이터 직접 결합 방식의 가이드 라인 추가
(안전한 비식별 데이터: 속성자(민감속성)들의 모든 부분집합이 m-유일성을 만족하도록 조치한 비식별 데이터)

다. 현 제도의 제약점

(1) 현재 k-익명성 프라이버시 모델의 재식별 가능성 존재

(가) k-익명성, l-다양성 모델의 한계

- k-익명성은 같은 준식별자 값을 갖는 동질그룹의 수가 최소 k개 이상 되도록 가공하여 준식별자를 미리 알고 시도되는 준식별자 기반 재식별 공격을 방어하는 목적으로 2002년 개발됨
- 단점 : 동질그룹의 민감속성값이 모두 같을 경우 민감정보가 유출될 수 있으므로 동질그룹의 민감 속성값을 l-개 이상으로 가공하는 l-다양성 추가 적용 제안됨
- k-익명성, l-다양성 모델만으로는 민감속성(일반 속성자) 값을 미리 알고 시도되는 민감속성 기반 공격을 막지 못함
(예:공격자가 신림동에 사는 37세의 여자인 김화복씨가 폐암에 걸린 사실을 미리 알고 있고 비식별 데이터에 30대, 여자, 신림동에 사는 5명의 동질 그룹에 폐암 걸린 사람이 오직 한사람일 경우 이 레코드가 김화복씨의 레코드임을 재식별함과 동시에 해당 레코드에 있는 김화복씨의 다른 민감속성 값을 알아낼 수 있음)

(나) 손실

- 개인정보 비식별화 조치 가이드라인은 빅데이터의 자유로운 유통을 통해 다양한 메시업 분석 활동을 위한 가이드 라인을 2016년 6월에 공표하였으나 k-익명성 모델의 불완전성으로 인해 데이터 유통을 시도하는 회사나 비식별화를 점검하는 전문기관에서 안정성을 평가하는 평가위원들이 현재의 가이드라인을 준수하여도 우리나라의 강력한 개인정보 보

호법을 어길 수 있다는 불안감을 갖게 되어 공표 이후 1년이 지났지만 비식별화 데이터의 유통이 거의 진행되고 있지 않고 있음.

- 또한 k-익명성의 불완전성은 개인정보 유출과 개인의 소유권을 중요시하는 시민 단체들의 강력한 반대로 빅데이터 활용 회사들의 비식별 데이터의 유통을 적극적으로 추진하지 못하고 있는 실정임.
- 이러한 이유로 여러 회사의 빅데이터를 결합하여 활용하는 데이터 기반 4차혁명 기술인 빅데이터, IoT, 인공지능 등의 국내 기술과 산업의 발전이 정체됨

(2) 임시대체키 생성을 통한 간접 데이터 결합의 문제점

(가) 임시 대체키 연계 모델의 한계

- 현행 가이드 라인에 따르면 두 회사의 데이터를 결합하기 위해서는 각 회사의 데이터를 개인이 식별되지 않게 비식별화를 수행한 이후, 이 두 비식별 데이터를 결합할 때는 개인별로 유일한 키값을 임시로 생성하는 임시 대체키를 사용하도록 되어 있음.
- 임시 대체키는 개인 1명에 대해서 유일한 킷값을 생성해서 데이터에 추가하는 것으로 개인 재식별이 다시 가능해 진다는 자기 모순적 과정을 반드시 수행해야 함.
- 데이터 결합 이후 모두 임시키를 회사와 전문기관에서 모두 삭제하도록 하고 있으나, 이로 인해 재식별 능력, 재식별 의도, 재식별 시도 불가 서약 등의 문서기반 서약을 통한 안정 장치만을 수행하여 행정적 과정을 증가시키고 재식별 가능성을 완전하게 배제하지 못함.

(나) 손실

- 전문기관을 거쳐야하는 현재의 간접 데이터 결합 방식은 전문기관에서 요구하는 행정적 업무를 추가적으로 수행해하므로 데이터 유통 회사 차원에서 시간적, 행정적, 비용적 부담을 요구하며 회사의 중요 데이터가 전문기관의 담당자와 평가자 등에게 공개된다는 부담을 갖게 되어 적극적이고 활발한 유통 의지를 가로 막고 있음
- 또한 데이터 유통은 판매자와 구매자간에 일어나는 상업 행위로 구매자는 데이터에 대한 비용을 지불했음에도 불구하고 자신의 데이터도 비식별해야 하고 전문기관으로 유출해야 하는 불합리한 과정을 반듯이 거쳐야 하는 단점을 가짐.
- 구매자는 판매자의 비식별 데이터를 자신의 원본데이터에 직접 결합한다면 자신의 데이터가 외부로 유출하지 않고 자신의 원본데이터와 판매자의 비식별 데이터와 직접 결합함으로 비식별화로 인한 데이터 손실을 최소화하여 빅데이터 분석 결과의 정확성을 높일 수 있음.

라. 제도 개선 방향

(1) 적정성 평가 개선 방향

○ 단기적 개선방향

구분	문제점	개선방안
가이드/ 안내서	가이드에 의료 정보만 예시되어 타 정보(금액, 시간 등)에 대한 비식별 조치 예시 부족	제조, 금융, 유통, 통신, 의료 등 산업별 가이드라인 제정 및 상세한 예시 제시(준식별자 및 민감정보의 구분, 민감정보에 대한 적절한 범주화 사례 등)
평가 기준	신청기관의 개량 분석 단계에서 적정 평가 기준값 설정 어려우며, 평가 단계에서 기준값에 따른 비식별 조치 반속	평가 접수 시 기초데이터 서면 제출을 통해 평가단의 서면검토 후 사전 평가 기준 가이드 안내
데이터 결합	신청서 등의 다수 양식 존재 및 방문 결합 절차 복잡	온라인 신청 접수 후 평가단 Pool 안내 등을 간소화된 표준 양식 배포, 절차 간소화 및 데이터 송수신 방안 등
	데이터 결합 전후의 적정정 평가의 중복, 소모적인 절차로 물리적 시간과다 소요	전문기관에서 데이터 결합 후 비식별화 수준을 평가하고 충분한 비식별 조치가 취해진 경우 적정성 평가 생략
	결합률 저하 발생	연계 정보(CI) 일부를 조합ㅎ는 등 결합율을 높일 수 있는 “임시 대체키”의 생성 가이드 및 사전 결합률 확인 절차에 대해서는 현재의 가이드에 추가

○ 장기적 개선방향

구분	문제점	개선방안
주기적/반복적 비식별화 및 결합	월간, 일간 등 주기적 비식별화 및 결합 활용 시 적정성 평가 반복	전문기관 핫라인 구성을 통한 실시간 결합 지원 체계 구축
		기술적 관리적 보호조치가 우수한 비식별 조치 기관에 대한 인증을 통해 외부 적정성 평가 생략

(2) 프라이버시 모델 개선(민감속석 기반 개인 재식별을 방지하는 m-유일성 모델의 추가)

(가) 속성자(Attribute value) 조치 기준 (개인정보 비식별 조치 가이드라인 6 페이지)

현행문구	✓속성자(Attribute value) 조치 기준 - 정보집합물에 포함된 속성자*도 데이터 이용 목적과 관련이 없는 경우에는 원칙적으로 삭제 * '속성자'란 개인과 관련된 정보로서 다른 정보와 쉽게 결합하는 경우 특정 개인을 알아볼 수도 있는 정보 - 데이터 이용 목적과 관련이 있는 속성자 중 식별요소가 있는 경우에는 가명처리, 총계 처리 등의 기법을 활용하여 비식별 조치 - 희귀병명, 희귀경력 등의 속성자는 구체적인 상황에 따라 개인 식별 가능성이 매우 높으므로 엄격한 비식별 조치 필요
수정문구	✓속성자(Attribute value) 조치 기준 - 정보집합물에 포함된 속성자*도 데이터 이용 목적과 관련이 없는 경우에는 원칙적으로 삭제 * '속성자'란 개인과 관련된 정보로서 다른 정보와 쉽게 결합하는 경우 특정 개인을 알아볼 수도 있는 정보 - 데이터 이용 목적과 관련이 있는 속성자 중 식별요소가 있는 경우에는 가명처리, 총계 처리 등의 기법을 활용하여 비식별 조치 - 희귀병명, 희귀경력 등의 속성자는 구체적인 상황에 따라 개인 식별 가능성이 매우 높으므로 m-유일성 모델 적용(총 v개의 민감속성자들이 존재할 경우 w개(< v)의 부분 집합에 대해 m-유일성을 보장하지 않는 레코드는 모두 삭제

(나) 현행 가이드라인 2-3 적정성 평가 단계 (개인정보 비식별 조치 가이드라인 9페이지)

현행문구	2-3) 적정성 평가 단계 : k-익명성 모델 활용 ✓적정성 평가 필요성 - 비식별 조치가 충분하지 않은 경우 공개 정보 등 다른 정보와의 결합, 다양한 추론 기법 등을 통해 개인이 식별될 우려 - 개인정보 보호책임자 책임 하에 외부전문가가 참여하는 「비식별 조치 적정성 평가단(이하, '평가단')」을 구성, 개인식별 가능성에 대한 엄격한 평가 필요 - 적정성 평가 시 프라이버시 보호 모델 중 k-익명성을 활용 - k-익명성은 최소한의 평가수단이며, 필요시 추가적인 평가모델(ℓ -다양성, t-근접성) 활용
수정문구	2-3) 적정성 평가 단계 : k-익명성 모델과 m-유일성 모델 활용 ✓적정성 평가 필요성 - 비식별 조치가 충분하지 않은 경우 공개 정보 등 다른 정보와의 결합, 다양한 추론 기법 등을 통해 개인이 식별될 우려 - 개인정보 보호책임자 책임 하에 외부전문가가 참여하는 「비식별 조치 적정성 평가단(이하, '평가단')」을 구성, 개인식별 가능성에 대한 엄격한 평가 필요 - 적정성 평가 시 프라이버시 보호 모델 중 준실별자들은 k-익명성을, 민감속성(일반속성자)들은 m-유일성을 활용 - k-익명성(l-다양성, t-근접성)은 준식별자를 통한 재식별 시도를 막고 m-유일성은 일반 속성자를 통한 재실별 시도를 막아 모든 속성값을 통한 개인의 재식별을 방어함

(다) 평가 수행 - 적정성 평가 기준값 결정 (개인정보 비식별 조치 가이드라인 13페이지)

현행문구	〈미국교육부 ...〉 - k=3은 안전도를 보장하는 최소한의 기준 - 5〈=k 〈=10은 안전도가 높은 수준 * k-익명성 값은 데이터의 제공을 합법적으로 허용하기 위해 제시된 기준
수정문구	〈미국교육부 ...〉 - k=3은 준식별자 기반 재식별 방어의 안전도를 보장하는 최소한의 기준 (5〈=k 〈=10은 준식별자 방어 안전도가 높은 수준) - 총 v개의 민감속성들에 대해 w=3개의 부분집합에 대해 m=3은 민감속성 기반 재식별 방어의 안전도를 보장하는 최소한의 기준 (5〈=v〈=10, 5〈=m〈=10은 민감속성 방어 안전도가 높은 수준) * k-익명성 및 m-유일성 값은 데이터의 제공을 합법적으로 허용하기 위해 제시된 기준

(라) 프라이버시 보호 모델

- 참고 3. 프라이버시 보호 모델에 "m-유일성 모델" 소개 문구 추가

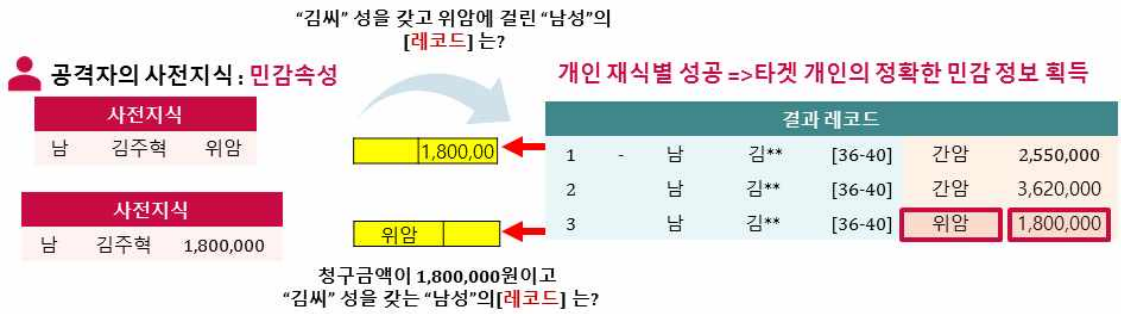
추가문구

(정의) m-유일성 프라이버시 모델
- 원본데이터의 민감속성 값을 기반으로 개인을 식별하는 시도를 원천적으로 방어하는 모델
- 원본데이터 셋 A와 이를 비식별화한 비식별 데이터 셋 B의 대응되는 동질그룹별로 데이터 셋 A와 B의 민감속성들이 총 v개 존재할 때, w〈=v)개의 부분 민감속성집합 값이 동일한 레코드가 존재하면 최소 m개 이상 존재하도록 가공함.

(예제)
- k-익명성, l-다양성 : 준식별자 속성의 연결 공격에 대한 방어 가능
- k-익명성, l-다양성 모델은 민감 속성을 통한 공격에 대한 방어가 불가능함
- 원인 : 민감속성에 대한 1-유일성 공격

	준식별자			민감속성	
구분	성별	이름	연령	병명	당월 청구 금액
1	남	김**	[36-40]	간암	2,555,000
2	남	김**	[36-40]	간암	3,620,000
3	남	김**	[36-40]	위암	1,800,000

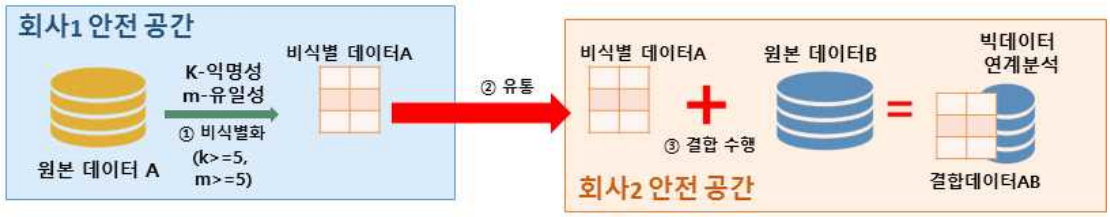
- 공격자가 민감속성에 대한 사전지식을 아래와 [그림 x]와 같이 갖고 있을 때 동질그룹에 해당 값이 오직 한 레코드에 존재할 경우 특정 개인이 재식별되도 다른 민감속성값이 유출됨 (k=3 k-익명성 보장 데이터)



	- 아래 표의 k=3 익명성, l=2 다양성 비식별 데이터도 민감속성기반 재식별 시도를 방어하지 못함. **구분 / 성별 / 이름 / 연령 / 병명 / 등급 / 당월 청구 금액(천원)** 1 / 남 / 김** / [36-40] / 간암 / 1등급 / 2,555 2 / 남 / 김** / [36-40] / 간암 / 1등급 / 3,620 3 / 남 / 김** / [36-40] / 위암 / 1등급 / 1,800 4 / 남 / 김** / [36-40] / **위암** / **2등급** / 1,800 5 / 남 / 김** / [36-40] / 위암 / 1등급 / 2,200 6 / 여 / 노** / [41-45] / **폐암** / 1등급 / 3,200 7 / 여 / 노** / [41-45] / 위암 / 2등급 / 1,670 8 / 여 / 노** / [41-45] / 위암 / 2등급 / 3,702 9 / 여 / 노** / [41-45] / 폐렴 / 1등급 / 2,500 - 공격자가 40대에 김씨 성을 갖고 위암 2등급인 남성이 이 데이터에 있다는 사실을 알고 재식별을 시도한다면 동질그룹(남, 김씨, 40-45세] 동질그룹에 부분 민감속성 〈병명과 등급〉의 값 조합을 한 레코드만 갖고 있음으로 1-유일성 공격으로 재식별 됨 **동질 그룹 의존 공격** 사전지식 남 김상호 40 위암 2등급 40세 "김씨" 성을 갖고 위암 2등급인 "남성"의 [레코드] 는? **개인 재식별 성공 =>타겟 개인의 정확한 민감 정보 획득** 4 남 김상호 40 위암 2등급 1,800

(3) 전문기관을 통하지 않는 직접 데이터 결합 방식 가이드

- 본 연구를 통해 임시 대체키를 생성하지 않고도 두 회사의 데이터를 결합할 수 있음을 실증하였음(SKT와 S신용평가 결합 실증)
- K>5이상이고 민감속성들의 모든 부분집합들에 대해서 m>=5이상을 만족하는 비식별데이터는 준식별자와 민감속성을 통한 재식별은 불가능함하고 임시대체키 자체를 생성하지 않는 결합 방법은 개인재식별의 위험성이 전혀 없음으로 전문기관을 통하지 않고 두 회사간에 자유로운 직접 데이터 결합을 아래와 같이 권장한다.
- 2-③ 절 이후 2-④절 새로 추가

추가문구	2-④ 별 데이터의 직접결합 방법 - 아래 조건을 충족시키는 비식별화 데이터의 결합은 전문기관을 통할 필요 없이 결합을 희망하는 두 회사 간 직접 데이터 결합 시행 (i) 결합 대상인 두 개의 데이터 셋 A와 B 중에 최소 한 데이터 셋(A)를 준식별자 속성에 대해 k-익명성(k)=5이상)을 충족시키면서 동시에 민감속성들의 모든 부분집합에 대해 m-유일성 (m)=5 이상)을 충족시키는 비식별화 데이터에 대해 (ii) 임시 대체키를 생성하지 않고 이 데이터 셋 A를 B의 원본 데이터나 비식별 데이터에 결합할 때는 전문기관의 적정성 평가 과정을 수행하지 않고 비식별화된 데이터를 다른 한 데이터 셋과 결합할 수 있다. (B는 원본데이터 또는 비식별 데이터이다.) **직접 결합 방식** 

제4장 목표 달성도 및 관련 분야 기여도

1. 목표 달성도

과제목표로 설정된 7개의 평가지표 달성 완료

세부 연구목표	세부 연구개발 내용	평가목표 및 착안점	진행경과 및 달성도
① 데이터 생성 및 가공	유통에 적합한 데이터 가공 비식별화 조치를위한 클린존 적용	* 거래용 비식별 조치 데이터 가공	• 가입자 미납 이력관련 데이터 1종 생성 및 비식별화 완료 • 사회적 약자 체류지/체류시간 데이터 1종 생성 및 비식별화 완료 • 외국인 가입자 체류지/체류시간 데이터 1종 생성 및 비식별화 완료
② 데이터 거래 실증	유통 플랫폼 구축 민간 공공영역 거래 실증	* 데이터 유통을 위한 플랫폼 환경 개발	• 빅데이터 허브 및 스마트인사이트 웹 환경 연계 개발 완료 • 웹 환경 내 3종 데이터 셋 게시완료 • SmartInsight연계 비정형 텍스트 데이터 자료 게시완료
③ KLT비식별화 알고리즘 실행	K L T 프라이버시 모델을 활용한 비식별 기술 적용 비식별 데이터 검증 기술 적용	* 재식별 불가성 * 비식별화 정확성	• 이지서티 솔루션 적용 비식별화 및 적정성 평가 완료 • 전문기관을 통한 한화 생명 데이터 연계 실증
④ MAS 비식별화 알고리즘 실증	저수준 MAS 비식별화 실증 고수준 MAS 비식별화 실증	* 재식별 불가성 * 비식별화 정확성 * 비식별 데이터 연계 활용성	• 그리즐리 솔루션 적용 비식별화 및 적정성 평가 완료 • 연계 활용성 검증을 위해 sci 평가 정보원 데이터 연계 실증
⑤ 비식별 결과 비교 분석	비식별 결과 비교 분석	* 비식별 기법들간 재식별 불가성, 정확성, 연계 활용성 비교 분석	• 연세대 데이터 베이스 연구실에서 비식별 기법간 지표분석 완료
⑥ 비식별 적정성 프로세스 실증	외부기관 인증 TTA / CC 인증	* 적정성 평가 프로세스실증 * 연계 정확성인증	• KISA, NIA 전문기관 적정성 평가수행 및 연계 실증 완료(2건/ 3회) • 참여기관 솔루션 TTA, CC인증 추진 완료
⑦ 홍보 및 교육사업	유통플랫폼 홍보 유통플랫폼 이용교육	* 과천과학관내 과제 결과 전시운영(1년간) 및 세미나 2회	• 과천과학관 내 AI 스피커를 활용한 빅데이터 비식별자료 전시 • 해외 전문가 세미나 2회 추진완료

2. 관련 분야 기여도

가. 데이터의 실제 유통 및 활용 가능성 실증

○ SK텔레콤 가입자기반 빅데이터를 국내 비식별화 기술 적용으로 재식별 가능성 없이 안전하게 유통 가능함을 실증

○ H보험과 S신용평가 데이터 연계사례를 통해 금융과 통신 비식별 데이터의 연계 가능성 실증

○ 특히 연계 데이터를 분석, 가공하여 중금리 대상자의 신용등급 구제가능성을 실증하여 신용등급관련 불리한 조건을 가진 서민계층 대상 SK텔레콤이 보유한 가입자 데이터를 연계하여 구제가 가능하다는 점을 실증하였음

○ 통신 가입자의 서비스 이용 데이터를 기반으로 민간, 공공 영역의 실수요 창출이 가능한 다양한 목적성 있는 데이터 가공이 가능하다는 시사점 확보

나. 신산업 창출 기대효과 시사점 확보

○ 외국인 가입자 및 사회적 약자의 체류지/시간 정보는 지자체 정책 수립 및 소상공인 활성화 기여 가능

○ 이동전화 수미납 이력 정보기반 데이터는 금융권의 중금리 대출 기준의 유연화를 통해 청년 창업 및 일자리관련 기회를 증대할 것으로 기대

○ 이종 산업간의 비식별 데이터 연계 실증을 통한 새로운 데이터 활용사례 및 서비스 창출 가능성 확보

제5장 연구개발성과의 활용계획

1. 개발 연구성과 활용 계획

활용계획	진행 방향
비식별 솔루션 체험관 제공을 통한 고객 접근성 및 사업화 루트 확보	- 프리웨어 버전을 웹사이트에 게시하여 관심 있는 고객사들이 체험할 수 있도록 안내 - 솔루션 소개 동영상 및 온라인 매뉴얼을 제공하여 비식별화에 대한 개념적 이해를 돕고 접근성을 강화
비식별 솔루션 기능 및 품질 향상에 따른 비식별화 프로세스 성숙화 도모	- 연구 측면에서의 비식별화 솔루션의 기능성 및 이론적 바탕을 견고하게 다듬을 수 있도록 산학간 연계 도모, R&D 투자 증액 계획 등 준비 - 반정형 및 비정형 데이터에 대한 비식별화 솔루션 개발을 완료하여 기능성 확장 및 범용성 확보 - 기 확보된 특허 및 인증 외의 지적재산권 추가 등록
기업 네트워크를 통한 개인정보 비식별화 관련 서비스 제공 및 추가 실증	- 본 사업 컨소시엄 내 참여기관들과의 지속적인 활용성 검토 및 통합 서비스 마련을 위한 논의 수행 (SKT 빅데이터 허브 등) - 비식별 조치 결합 기술에 대한 추가적인 실증 사례 확보
빅데이터 분석 활성화 및 관련 서비스 고도화 실현	- 개인정보 비식별화를 통한 개인정보 보호의 문제가 실증적으로 해결됨으로써 빅데이터 분석 활성화 환경 조성을 위한 가이드 역할 수행 - 유사한 특징 및 성향을 가지는 개인들의 관심에 맞춘 차별화된 서비스가 가능해지므로 유관 업체와의 관련 서비스 구축 협력

가. 비식별 조치기술 및 유통 플랫폼의 활용방안

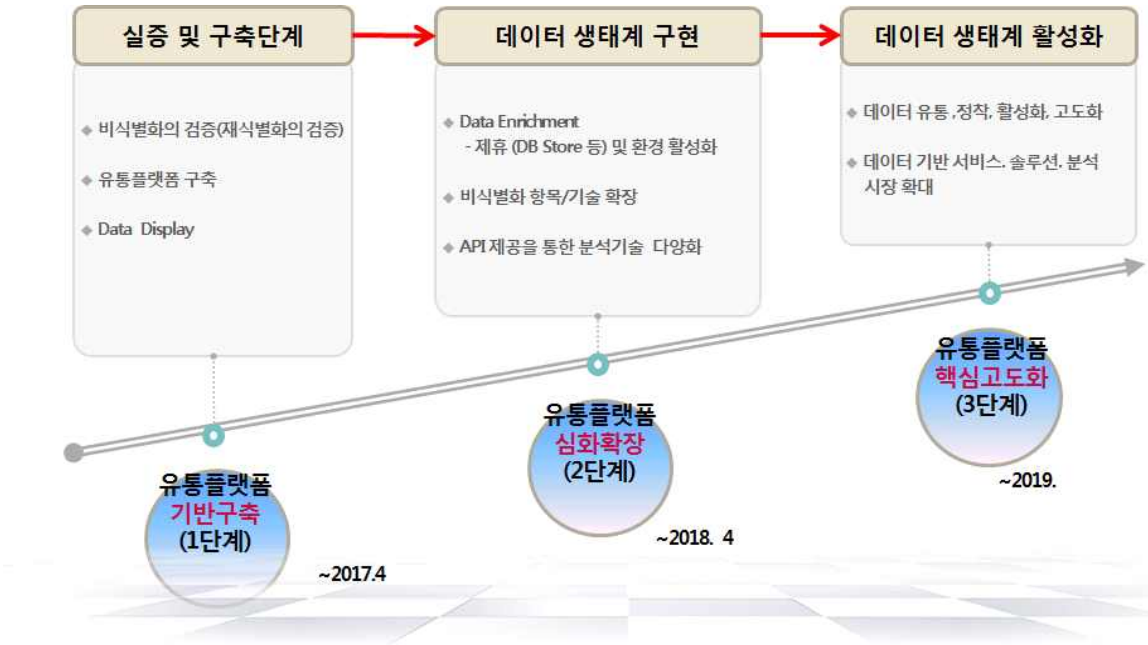
○ 빅데이터에서의 개인정보 비식별조치를 통한 빅데이터 유통 활성화로 인한 시장 확대
- 개인정보 비식별 조치를 통해 프라이버시 문제 해결, 공공 및 민간 빅데이터 유통 및 개방

○ 4차 산업혁명 시대를 맞아 빅데이터 공유, 거래에 적합한 데이터 활용 플랫폼 확보로 다양한 관련 산업 분야에서 지속적인 데이터 공급 추진

○ 산업 분야별 맞춤형 비식별화 데이터 제공, 연계 분석 컨설팅 등이 가능한 데이터 생태계 구축 기반으로 활용

○ 적절한 비식별화 프로세스를 거친 데이터의 재식별 불가성에 대한 실증 내용을 바탕으로 향후에도 홍보활동을 병행하여 대국민 프라이버시 리스크 인식 개선활동 추진예정

○ 중, 장기적인 단계별 데이터 유통사업 활성화 방안을 가지고 시장 개척을 추진하며 유통 플랫폼 활성화를 위한 지속적 투자

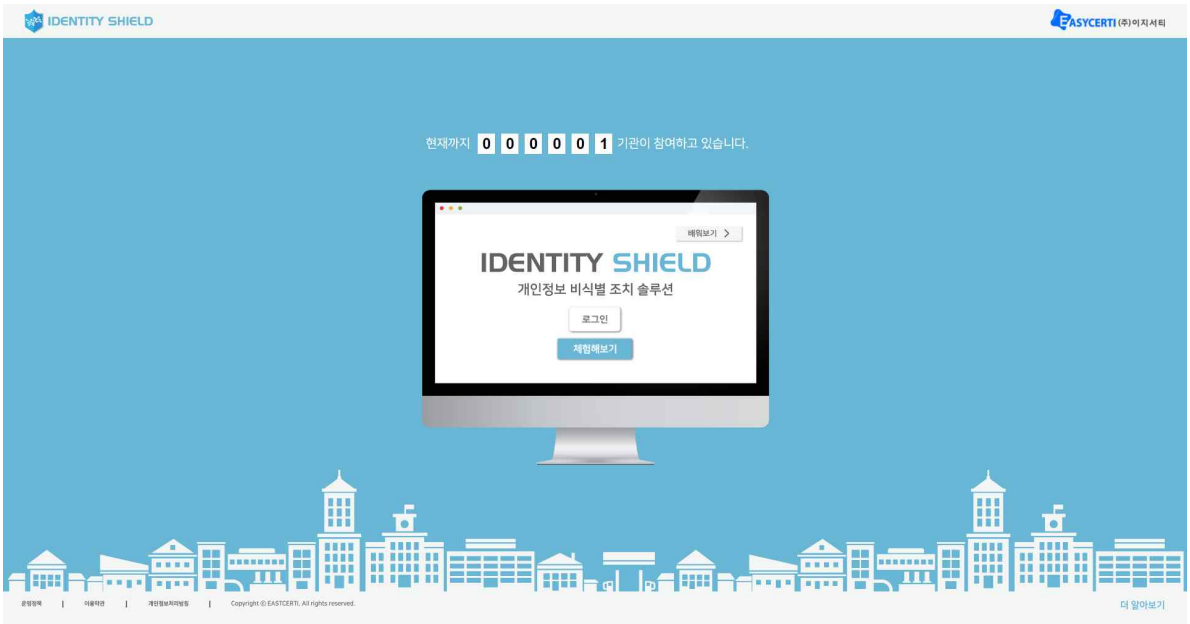


2. 비식별 조치 시범 서비스

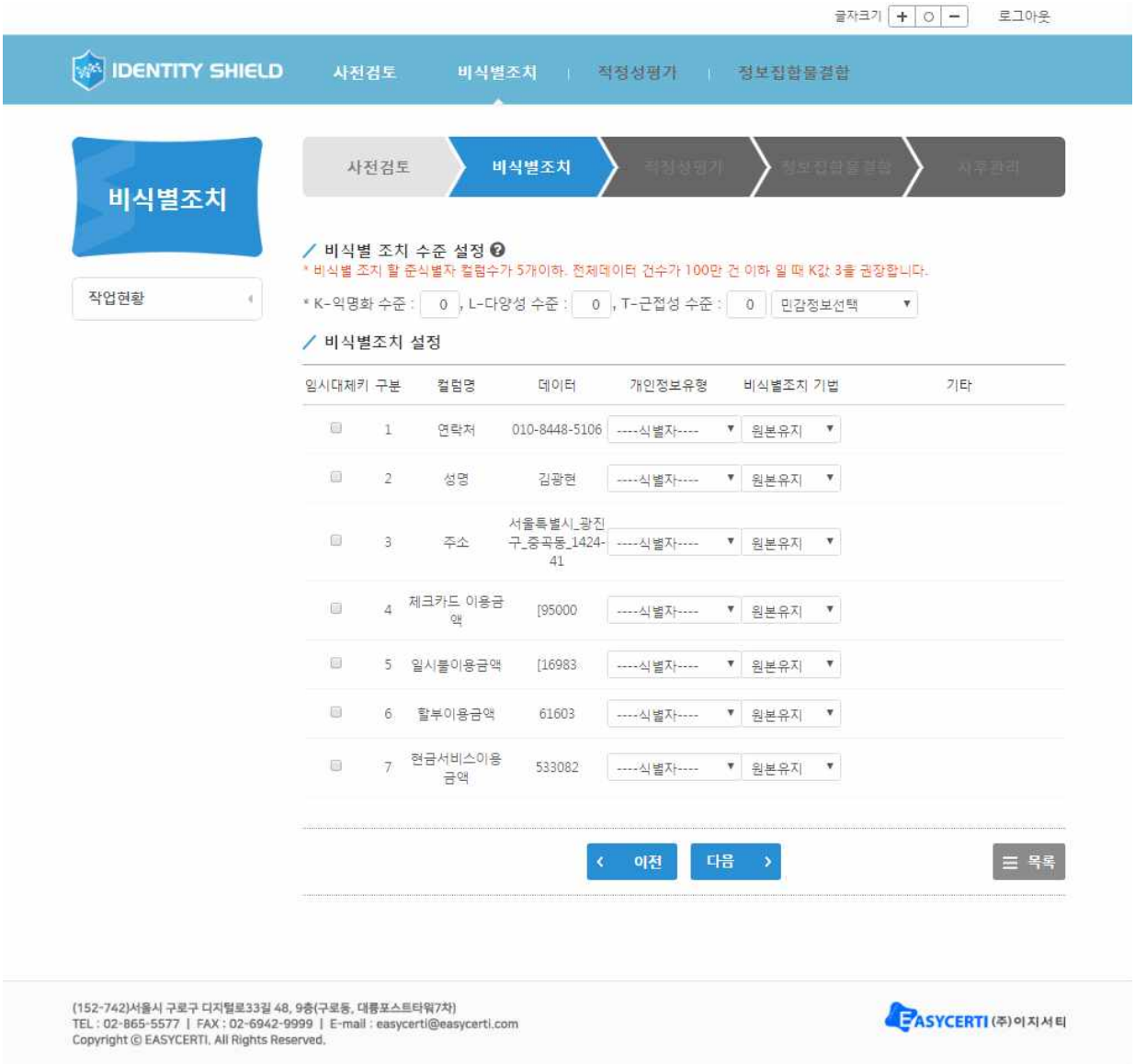
가. IDENTITY SHIELD 시범서비스

(1) IDENTITY 시범서비스 오픈

○ 주식회사 이지서티의 IDENTITY SHIELD(개인정보 비식별 조치 솔루션) 시범서비스를 IDC 센터를 이용하여 자체적으로 오픈하였음



[그림 5-1] IDENTITY SHIELD 시범서비스 화면

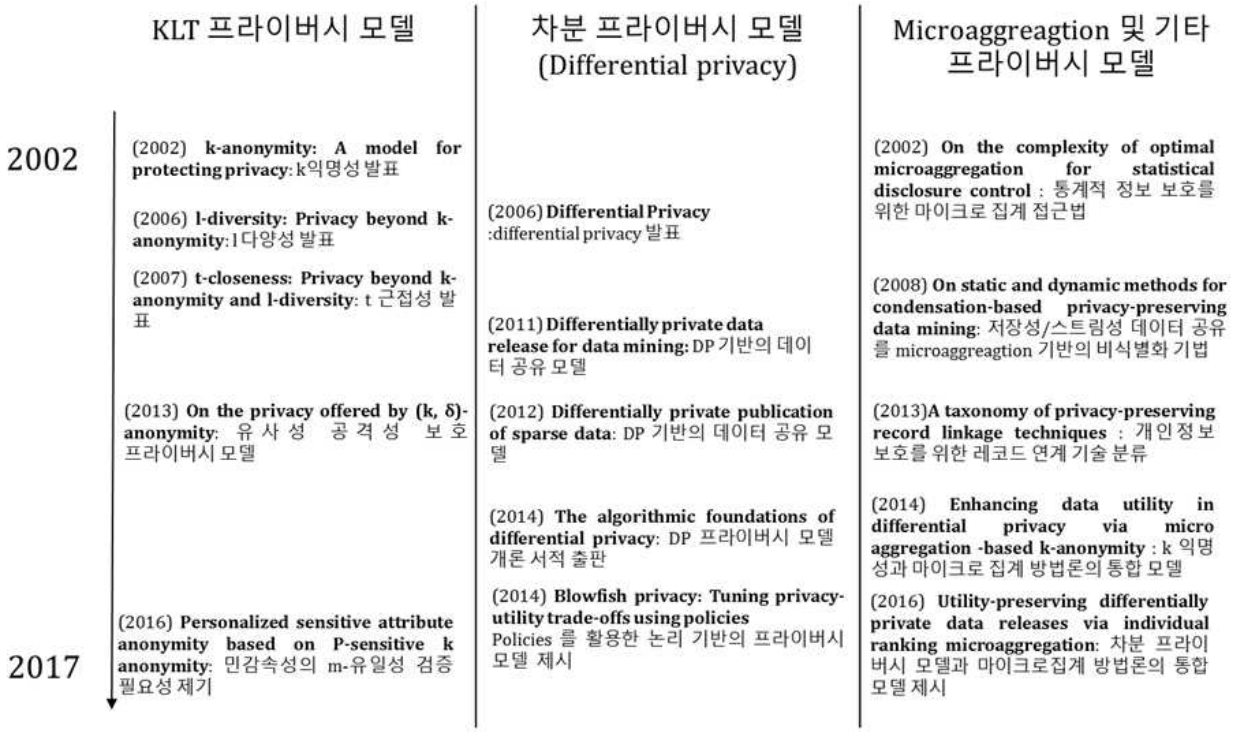


[그림 5-2] IDENTITY SHIELD 비식별 조치 설정 화면

제6장 연구 과정에서 수집한 해외 과학기술 정보

1. 기술현황

가. 프라이버시 모델



(1) k-익명성, l-다양성, t-근접성 모델

○ Very Large DataBase(VLDB), IEEE Transactions on Knowledge and Data Engineering (TKDE), ACM Transactions on Database Systems (TODS) 등의 데이터베이스, 데이터 마이닝 분야의 해외 SCI 급 저널 및 학술대회에서 일반적으로 Privacy preserving Data Publication(PPDP)이라는 키워드로 매년 활발히 연구 실적들이 발표 중

○ k-익명성, l-다양성, t-근접성을 모두 적용한 데이터 세트에 대해서도 특정한 사전지식이 있는 경우 최소성 공격 등을 통해 개인정보가 유출될 수 있음 증명됨

○ 일반적으로 준식별자, 민감속성으로 확실히 역할을 나눌 경우 민감속성을 통한 연결 공격에 대한 제기되고 이에 대한 해결알고리즘이 제시됨

○ 비식별레코드 세트에 대한 유사성 연계 공격에 대한 위험성이 제기 되었으며 (k, δ)-anonymity모델을 통해 이를 보호

○ 개인정보는 보호하되 개인을 정확히 연계하기 위한 기술 연구 또한 진행 중이며, Privacy-preserving record linkage (PPRL)이라는 키워드로 성과가 발표 중

✓ 연결 공격: 공격자가 알고있는 사전정보를 이용하여 비식별화 결과 레코드 세트의 속성값과의 연계를 통해 대상 레코드나 원본 속성값을 유추하는 공격

(2) 마이크로집계 모델

○ VLDB, TODS, Information Fusion 등의 데이터베이스, 데이터 마이닝 분야의 해외 SCI 급 저널 및 학술대회에서 PPDP 뿐만 아니라 PPDM (Privacy Preserving Data Mining) 분야에서도 관련 연구가 활발히 진행 중

○ 최소 레코드 수 k의 최적화/ 레코드 파티셔닝에 대한 최적화/ 속성값 집계처리의 방법론 등 다양한 관점에서의 연구가 수행됨

○ 저장성 데이터뿐만 아니라 스트림 데이터에 대한 레코드 공유에 대한 연구도 이뤄짐

○ k-익명성 프라이버시 모델과의 통합된 프라이버시 모델에 대한 연구가 이뤄짐

○ 차분 프라이버시 모델과 마이크로집계 모델의 통합 프라이버시 모델에 대한 연구가 진행됨

(3) 차분 프라이버시 모델

○ ACM SIGKDD Conferences on Knowledge Discovery and Data Mining, Information Sciences, ACM SIGSAC conference on Computer & communications security 등의 다양한 분야의 SCI 급 저널 및 학술대회에서 PPDP 뿐만 아라 니PPDM (Privacy Preserving Data Mining)분야에서도 관련 연구가 활발히 진행 중

○ 데이터베이스, 머신러닝, 보안 등의 데이터 공유를 뿐만 아니라 데이터 분석을 위한 대부분의 기술에 프라이버시를 보호하는 기본 모델로 적용 중

○ 일반적으로 통계 데이터베이스 기반의 분석 모델에서 적용 가능한 프라이버시 모델이지만 다양한 변형된 기법들이 제안되면서 데이터 공유를 위한 방법론으로 변형되어 사용되기도 함

○ 노이즈 추가 방식에 있어서도 Laplace mechanism 외에 exponential mechanism, net-mechanism, SuLQ mechanism 등의 다양한 기법들이 제안됨

○ 차분프라이버시 모델은 수리 통계적 모델링을 통해 개인정보 유출에 대한 문제를 확률적으로 해당 모델의 발전 모델로 논리기반의 프라이버시 모델이 제안됨

나. 프라이버시 보존 데이터 추출(Privacy Preserving Data Mining) 처리 기술

(1) 국외 학계를 중심으로 PPDM에 관련된 연구가 활발히 진행 중이며 PPDM의 기술적인 구분은 크게 Anonymization, Perturbation, Randomized Response, Condensation approach, Cryptography 기반의 접근법으로 나눌 수 있음

(2) 국외에서는 최근 들어 일부 의료계나 방법 시스템에 특화되어 사용되고 있던 PPDM 기술의 범위가 실제다양한 도메인의 데이터에 적용할 수 있도록 지원하는 제품들이 공개 중

○ 버클리에서 2012년 PPDM을 위한 오픈소스 플랫폼 GUPT 시스템을 제작 및 공개함

다. 정보손실 방지를 위한 통계적 정확성 유지 비식별화 기술

(1) (국외) 글로벌 DB 벤더 회사들은 분석시스템 내에 비식별화 기능이 추가된 솔루션이나 비식별화 기능을 제공하는 제품을 제공하며 IBM을 선두로 하여 실시간 빅데이터 비식별화 기술의 연구 및 상용화가 이루어지고 있음

○ IBM에서 제공하는 InfoSphere Data Privacy 솔루션에서 데이터 스트림의 실시간 암호화, 비식별화 변환을 지원

○ IRI에서 제공하는 fieldShiled 제품은 SQL 로직을 활용한 반 실시간성의 비식별화 기술이 포함됨

(2) (국외) 향후 데이터 분석 시 품질유지를 위한 비식별화 기술이 학계에서 다양하게 제보되고 있으며, 업계에서도 이러한 문제를 해결하기 위한 기술 적용의 필요성이 증대됨에 따라 일부 제품이 공개된 상황

○ The International Household Survey Network (IHSN)는 microdata 제공 및 사용을 원하는 사람들을 위한 R에서 사용가능한 SDCMicro 오픈소스 패키지를 공개

라. Hewlett Packard Enterprise의 SecureData

(1) 특징

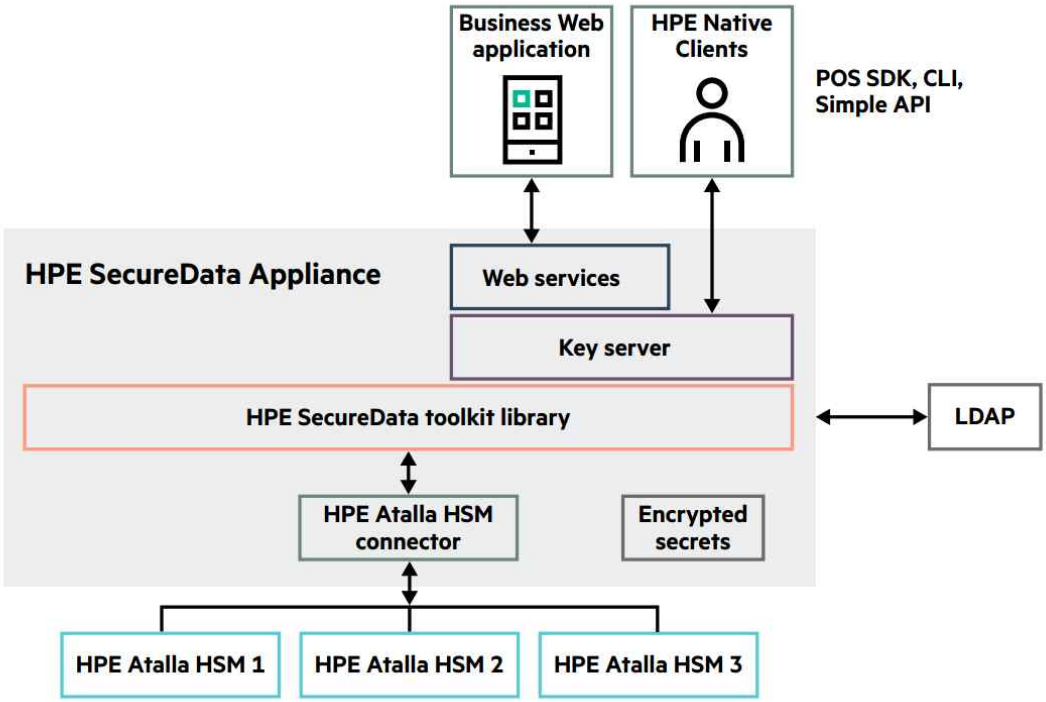
○ 개인정보를 토큰화하여 개인정보 공격을 방지하는 솔루션

○ 유럽 연합을 위한 GDPR 법안의 암호화 및 가명지침을 지원

○ 중앙 정책 관리 및 제어

○ 단순하고 고성능의 네이티브 플랫폼 API

(2) 구성도

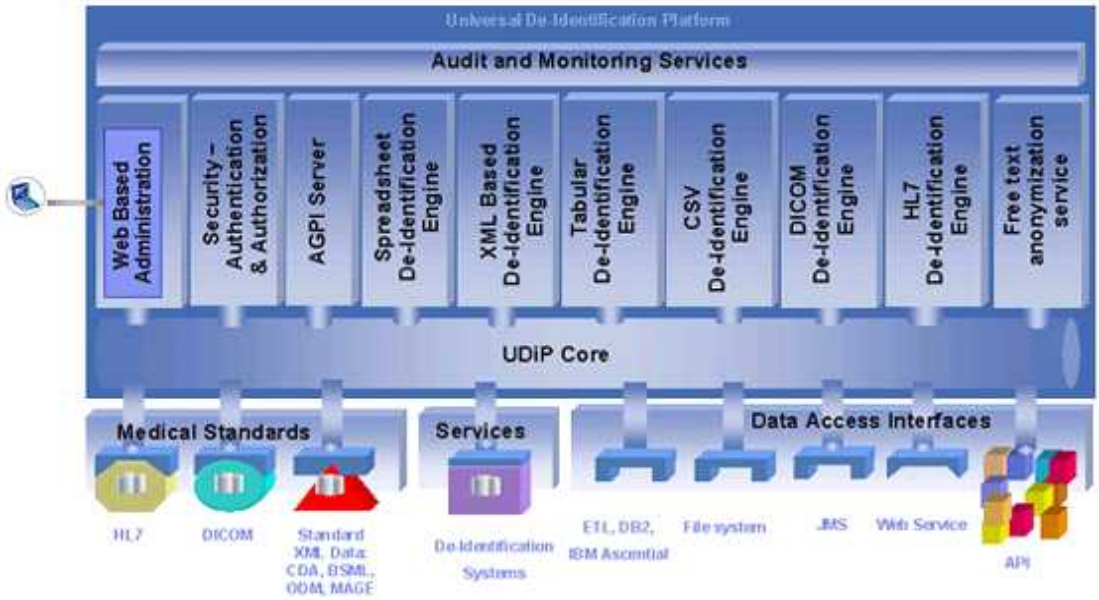


마. IBM의 UDiP(Universal De-identification Platform)

(1) 특징

- 의료 데이터에 대한 개인정보 비식별화를 수행하여 그 결과 레코드들로부터 추가적인 분석 정보를 추출하는 프레임워크
- 플랫폼으로 구현되어 내부 모듈들이 상황에 따라 플러그인/아웃을 할 수 있음
- XML 문서, CSV 및 스프레드시트 파일, 자유 형식의 텍스트 파일 등 다양한 형태의 반정형/비정형 데이터를 지원
- 자바 EJBean 기반의 API 형태 및 웹 서비스, IBM 보유 WebSphere 솔루션 등 여러 방식의 인터페이스를 지원
- HIPAA 및 HL7, DICOM, CDA, BSML, ODM, and MAGE 등의 의학 표준을 따르고 있어서 의료 데이터 도메인에서는 매우 높은 활용성 지님

(2) 구성도



Universal De-identification Platform (UDiP)

2. 국내 외 법 제도 정비 연구

가. 캐나다의 구조적 데이터에 대한 비식별화 조치 가이드라인 검토 (De-identification Guidelines for structured data, June 2016)

가이드라인에 따르면 “비식별화 (De-identification)”는 일반적으로 데이터 세트 또는 기록에서 개인정보를 제거하는 절차를 의미하는 용어이다. 비식별화는 개인의 프라이버시를 보호하는 것으로 비식별화를 하면 데이터 세트에 더 이상 개인 정보가 포함되어 있지 않기 때문이다. 만약 데이터 세트에 개인 정보가 들어있지 않으면 그것을 사용하거나 공개해도 개인의 프라이버시가 침해되지 않는다. 그렇게 되면 정보공개 및 개인정보보호법 (Freedom of Information and Protection of Privacy Act; FIPPA) 및 지방자치체 정보의 자유 및 개인정보 보호법 (Municipal Freedom of Information Protection of Privacy Act; MFIPPA)은 비식별화된 정보에는 적용되지 않는다라고 명기되어 있다.

또한 가이드라인에서는 공개된 데이터의 재식별화 위험성의 허용수준을 산출하는 것이 포함되어 있는데 위험도를 계산할 때 데이터 세트 내에 타겟 개체가 포함되어 있는지 또한 이를 변호사가 알아낼 수 있는지를 “검사 (prosecutor) 위험”이라 정의하고 있다.

공개 데이터의 가치에 대해서는 온타리오의“열린 데이터” 계획이 데이터 세트를 능동적으로 공개 하고 모두가 자유롭게 사용하고 재게시할 수 있도록 만드는 것을 통해 정부의 투명성 및 책임을 향상시키도록 추구하고 있으며 이런 계획들이 제공하는 정보의 유용성 및 증가된 양을 고려해보면 기관들이 개인의 프라이버시를 보호하는 방법으로 자신의 데이터 세트를 공개하는 것이 중요하다고 밝히고 있다.

나. 캐나다 가이드라인의 구조적 데이터의 비식별화 절차

정보 내에 있는 실용성을 가능한 한 보존하면서 개인의 프라이버시를 보호하기 위해서는 데이터 세트의 공개와 관련된 재식별화 위험성의 종류 및 정도의 구조적 분석을 통해 비식별화의 종류 및 양을 결정해야 한다. 데이터 세트를 비식별화 하는 것을 시도할 때 다음의 절차를 고려해야 한다.

(1) 공개 모형을 결정한다.

데이터를 게시할 때에는 누가 정보에 접속하고 어떻게 접속하는지를 포함해서 제한을 가능한 최소로 하는 것이 일반적인 관행이다. 데이터를 게시하는 기관들에 자신을 등록 및 식별하게 하는 개인에 대한 요구사항들은 접근, 사용 및 개인이 정보를 찾을 수 있게 하는 것을 막는 장애물로 고려된다. 이와 같이 데이터 세트를 다운로드 받은 개인들이 식별되지 않는다면 이 같은 공개는 공개적 데이터 공개와 같은 경우로 다루어져야 한다.

하지만 개인을 등록하고 신원을 확인하는 것을 요구할 때도 있다. 예를 들면, 정부 또는 대학에서 협찬한 프로그래밍 대회 또는 “해커톤”에서 비식별화된 데이터 세트를 대중 또는 전학생에게 공개할 수도 있는데 특정 방법으로 (데이터 세트 안에 있는 개인을 재식별화하거나 사용조건 동의서를 통해 제 삼자에게 정보를 공개하는 것을 포함) 참가자들이 데이터 세트를 사용하지 못하도록 제한할 수도 있다. 사용조건 계약서에 참가자들에게 추가적인 프라이버시 및 보안 조치를 요구하지 않는다거나 또는 그런 조치들이 강요하지 않는다면 이와 같은 공개는

반공개적 데이터 공개로 다루어야 한다.

마지막으로, 기관 간에 정보를 공유할 때는 데이터 세트 접근이 수령 프로그램 영역 또는 기관으로 제한되어 있기 때문에 정보의 프라이버시 및 보안은 데이터 공유 계약서를 통해 설정하고 시행할 수 있다. 이 경우에는 그런 공개가 비공개적 데이터 공개로 다루어질 수 있다.
데이터 공개가 비공개적인 것으로 고려되기 위해서는 각 당사자간에 데이터 공유 계약서가 있어야 한다. 이러한 공개들의 경우에는 데이터 공유 계약서가 위험 완화 전략의 중요한 부분을 차지한다.

(2) 변수를 분류한다.

만일 데이터 세트가 개인에 관한 것이라면 한 파일에서 각 행은 한 개인을 나타내고 각 열은 개인으로부터 수집된 정보의 변수들을 나타낸다. 정보의 종류에 따라 어떤 변수들은 직접적으로나 간접적으로 개인을 식별하는데 사용될 수 있지만 어떤 것들은 그렇지 않을 수 있다. 비식별화는 개인을 식별하는데 사용되는 변수들만 고려한다. 위에서 언급했듯이 두 종류의 변수들이 있는데 이는 직접적 식별자와 간접적 식별자 또는 준식별자이다.

(3) 재식별 위험 한계치를 결정한다.

비식별화는 개인을 식별하는 정보 또는 개인을 식별하기 위해 하나 단독으로 또는 다른 정보와 같이 사용될 수 있다는 타당한 근거를 가진 정보를 제거하는 것을 통해 개인의 프라이버시를 보호한다. 개인의 프라이버시를 보호하기 위해서 적용되어야 하는 비식별화의 정도는 데이터 세트를 공개한 경우 재식별화 위험의 정도와 비례한다. 데이터 공개의 재식별화 위험도가 높을수록 더 많은 정도의 비식별화가 요구된다.

잠재적인 개인의 프라이버시 침해의 정도를 평가할 때는 데이터 세트에 있는 정보가 식별가능하고 비식별화 작업이 이루어지지 않았음을 가정해야 한다. 이 가정하에 프라이버시의 침해 정도는 다음을 포함한 여러 다른 요인들의 함수이다:

- 정보의 민감성
- 정보의 상세함의 정도 및/또는 범위
- 개인들의 수
- 부적절한 사용 또는 위반의 경우에서 개인에게 가해지는 잠재적인 부상 또는 피해
- 개인들의 동의 없이 FIPPA 또는 MFIPPA에 따라 정보의 공개가 허가 되었는지
- 프라이버시에 대한 예상 적거나 아예 없는 채로 개인들이 거리낌없이 정보를 주었는지 또는 임의의 정보
- 개인이 이와 같은 이차적인 목적을 위해서 비식별화된 형태로 정보가 공개되는 것에 분명하게 동의를 했고/또는 데이터를 수집할 당시 이에 대해 통지를 받았는지[11]

프라이버시 침해의 평가 결과는 질적 가치이지만, 데이터 세트에 적용되어야 하는 비식별화의 정도는 절대치로 수량화해야 한다. 이 차이를 줄이기 위해서는 프라이버시 침해 값을 평가한 후 그 결과를 그 위험 정도와 비례하는 비식별화 정도를 나타내는 숫자 값으로 변환해야 한다. 이 “재식별화 위험 한계치”는 일반적으로 비식별화 되었다고 고려되는 정도 즉, 더 이상 개인정보가 포함되지 않도록 하는 최소한의 비식별화 작업을 데이터 세트에 적용해야 한다. 그에 맞춰 이는 비식별화와 관련된 당신의 산출을 비교할 수 있게 하는 기준치를 형성해준다.

프라이버시 침해의 (질적) 값과 (양적) 재식별화 위험 한계치간 전환할 때 비식별화의 주요 측면인, 비식별화를 통해 재식별화 확률이 제로인 데이터 세트를 생성할 수 없다는 것을 고려해야 한다. 그 보다는 공개될 때 관련된 재식별화의 위험 정도를 고려해보면 재식별화의 확률이 매우 낮은 데이터 세트를 낳는다. 프라이버시 침해 값에 비례하는 비식별화의 정도는 그 위험의 정도를 고려했을 때 매우 낮은 재식별화 확률과 같아야 한다.

다음 표는 여러 다른 프라이버시의 침해 값을 가진 데이터 세트에 대한 매우 낮은 재식별화의 확률로 여겨지는 것이 무엇인지를 결정하는데 사용할 수 있는 지표이다[12].

프라이버시 침해	재식별 위험 한계치	동등한 셀 크기
하	0.1	10
중	0.075	15
상	0.05	20

(4) 데이터 위험을 측정한다.

허용되는 재식별화 위험 한계치를 결정한 후 다음 단계는 데이터 세트 내에 있는 재식별화 위험의 정도를 측정하는 것이다. 데이터 위험은 공개와 관련된 재식별화의 정도를 결정하기 위해서 사용된다.데이터 세트에 있는 재식별화 위험의 정도를 측정하는 것은 두 단계 절차이다. (1)각 행의 재식별화의 확률을 산출하고 (2) 사용하는 공개 모형에 따라 적합한 위험 측정 방법을 적용한다.

개인들에 대한 데이터 세트의 각 행에는 한 개인에 대한 정보가 들어있다. 따라서 각 행은 재식별화의 확률을 가지고 있다. 한 행의 경우 재식별화의 확률은 데이터 세트 내에 준 식별자인 변수와 동일한 값을 가진 다른 행들이 몇 개 인지에 따라 달려있다.

준 식별자인 동일한 값을 가진 변수들을 가진 데이터 세트의 모든 행은 “등가류”를 형성한다. 예를 들면, 성별, 나이 및 최종학력의 변수들을 가진 데이터 세트에서 중등과정 이후의 학위를 가진 35세 남성에 해당하는 모든 행은 등가류를 형성한다. 등가류의 크기는 준 식별자에 대한 동일한 값을 가진 행의 개수와 동일하다.

각 행의 경우 재식별화의 확률은 1나누기 그 등가류의 크기이다. 예를 들면 등가류 크기가 5인 경우를 가진 각 행은 재식별화의 확률이 0.2이다.

$$\text{주어진 행의 재식별화 확률} = 1/(\text{등가류크기})$$

(5) 문맥적 위험을 측정한다.

데이터 세트에서의 위험이 데이터 세트의 공개와 관련된 재식별화 위험의 정도를 결정하는 중요한 역할을 하지만 이 요인만이 고려할 대상이 아니다. 재식별화 위험은 주어진 사용한 공개 모형에서 데이터 세트에 가해질 수 있는 재식별화 공격의 여러 종류의 함수이기도 하다. 가능한 공격에 관해서 재식별화 위험을 추가로 분석하는 것은 문맥적 위험을 가져온다. 데이터 위험과 함께 이 값은 데이터 세트의 공개와 관련된 재식별화의 총 위험을 산출하는데 사용된다

문맥적 위험은 데이터 세트에 대해 하나 또는 그 이상의 재식별화 공격의 확률이다. 비식별화된 데이터 세트가 공개되자마자 재식별화 공격이 가해질 수 있지만 공격의 종류는 사용하는 공개 모형에 따라 달라진다.

(6) 총 위험을 산출한다.

데이터 위험 및 문맥적 위험이 측정된 후에는 총 위험을 산출할 수 있다. 총 위험은 데이터 위험에 문맥적 위험을 곱한 값이다.

총 위험= 데이터 위험 x 문맥적 위험

총 위험은 공격이 가동되었을 경우 하나 또는 그 이상의 행이 재식별 되는 확률과 동등하다. 예를 들면 한 데이터 세트의 데이터 리스크가 0.2이고 문맥적 리스크가 0.5일 경우 그 데이터 세트의 총 위험은 0.1이다.

(7) 데이터를 비식별화 한다.

데이터 세트가 비식별화 된 것으로 고려되려면 식별 가능한 모든 정보가 제거되어야 한다. 개인을 식별하는 모든 정보 또는 개인을 식별하기 위해서 단독적으로 또는 다른 정보와 함께 사용될 수 있다고 합당하게 추론되는 정보를 제거하기 위해서는 데이터 세트의 값이 여러 방법으로 변형될 수 있다. 식별자의 종류 및 특성에 따라 다른 기술들을 적용할 수 있다. 식별 가능한 정보를 제거하기 위해서는 다음을 행해야 한다:

- 직접적 식별자를 마스크 한다.
- 등가류의 사이즈를 변형한다.
- 총 위험이 재식별 위험 한계치 이하가 되도록 한다.

(8) 데이터 실용성을 평가한다.

데이터 세트에 적용된 비식별화의 정도와 발생된 정보의 유용성간에 상충효과가 있을 수 있다. 준식별자라 여겨지는 변수들이 일반화 또는 억제와 같은 기술들을 사용해 비식별화 될수록 이에 해당하는 데이터 세트의 유용성에 손실이 커진다.

재식별의 총 위험이 재식별 위험 한계치 이하가 되도록 데이터 세트에 일반화(generalization)와 억제(suppression)가 적용되지만 이 결과를 얻기 위해서는 이 비식별화 기술들이 여러 다른 방법 및 조합으로 적용되어야 할 수도 있다. 예를 들어 한 접근법은 등가류의 크기를 증가시키기 위해 일반화 및 범주의 정확성을 감소시키는 것에 중점을 두고 있을 수도 있다. 또 다른 접근법은 억제 및 등가류가 너무 작은 변수의 셀 또는 열을 제거하는 것에 중점을 두고 있을 수도 있다. 데이터 세트의 특성에 따라 다른 적용 및/또는 일반화와 억제의 조합이 개인의 프라이버시를 보호하면서 정보에 있는 유용성을 더 보존할 수 있다.

일반적인 규칙으로는 데이터 세트에 있는 열의 5 퍼센트 이상이 이미 어느 형태의 억제가 적용되어 있지 않은 경우 일반화보다 억제(suppression)를 먼저 고려해야 한다. 일반화가 데이터 세트 내의 모든 열의 정확도를 감소시키는 반면 억제는 한 열의 정보를 제거하기 때문에 비식별화를 위해서는 억제를 시작점으로 고려할 수도 있다.

일반화와 억제의 기술을 새로운 방법으로 적용 및/또는 조합하면 재식별의 총 위험이 위험 한계치 이하에 있도록 하면 더 높은 유용성을 가진 데이터 세트를 생성할 수 있다.

(9) 과정을 기록한다.

개인 정보가 들어있는 데이터 세트를 비식별하는 각 시도는 동일한 단계를 따르고 동일한 문제들을 평가해야 한다. 하지만, 변수들과 값, 그리고 비식별화의 종류 및 양을 결정하는 분석

은 각 데이터 공개마다 다를 것이다. 개인 정보를 비식별화 작업을 하는 과정과 관련된 복잡함과 도전들에 대한 안내를 원하면 절차 및 결과를 기록하는 보고서를 작성하는 것을 고려해야 한다. 이 우수 관행에 대한 다음을 포함한 여러 이점들이 있다

3. 국. 내외 시장연구

가. 글로벌 데이터 거래관련 현황

○ 시장조사업체 오범(Ovum)은 전 세계 빅데이터 시장이 2016년 17억 달러에서 2020년에는 94억 달러까지 성장할 것으로 전망.

○ 해외시장(특히 미국)은 다양한 데이터 거래 사업자가 시장을 형성 중

○ 해외 데이터브로커, 또는 정보 재판매업자(information reseller)의 경우 브로커간의 데이터 유통도 매우 활발하며 시장이 확대되고 있는 상황(미국 주요 9개 데이터 브로커 매출은 ’12년 기준 426.7백만달러)

〈표 6-1〉 주요 데이터 사업자 현황

- 마케팅캠페인, 부정사용 탐지를 위한 고객데이터 분석 서비스 제공
- l전 세계 7억 명의 소비자 정보가 담긴 데이터베이스 보유
- 산업계와 정부에 재무정보와 부동산정보에 기초한 분석서비스 제공
- 약 8억 건의 부동산 거래정보, 약 1억 건의 담보 데이터베이스 보유
- 거의 모든 미국 소비자의 마케팅 데이터를 제공
- 2012년 페이스북은 페이스북 이용자의 소셜사이트 상품광고 조회와 오프라인
- 상점의 구매 관련성 측정 위해 데이터로직스와 협력 발표
- 마케터와 재무관련 회사, 온라인유통업체에 수익성이 높은 잠재 고객과 부정 거래 예측 서비스 제공
- 매달 평균 30억 건이 넘는 새로운 정보 추가 축적
- 특정인 확인, 부정 거래 확인 서비스 제공
- 7천억 건의 데이터와 14억 건의 소비자 거래 데이터 보유
- 신원 조회와 공문서 정보 제공
- 20억 건이 넘는 데이터베이스 보유
- 소셜미디어사이트, 홈페이지, 블로그의 콘텐츠를 분석 작성자 확인 서비스 제공
- 소비자와 기업의 과거 이력 데이터 분석을 통해 미래 행동 예측 정보 제공

※ 2013년 10월 기준으로 총사업자 수 약 650개, 2012년 동안 약 1,560억 달러의 추가 매출이익이 발생했고, 고용인원은 약 67.5만명으로 추정됨
위 사업자 매출계는 4억 3천만달러임/ 출처 : 정용찬, “빅데이터 산업과 데이터 브로커”, 정보통신정책연구원, 2015

나. 국내 데이터 산업 현황 (참조: SPRi)

○ 국내 기업의 빅데이터 도입율은 전체기업 기준으로 약 4.3% 수준

○ 국내 기업의 빅데이터 활용 성장률은 중대형 업체의 경우 20~25%, 중소업체의 경우 5~8% 수준

○ 국내 기업의 향후 빅데이터 수요는 전체기업의 30.2% 수준이며, 도입 고려 시기는 2018년(77개)과 2019년(98개) 이후가 많음

○ 국내의 빅데이터 기술 수준은 선진국(100) 대비 62.6% 수준이며 기술 수준 격차는 약 3.3년 뒤쳐져 있는 것으로 분석

○ 국내 기업들의 빅데이터 분석 도입 수준이 뒤처지는 이유는 빅데이터 분석을 할만큼 풍부한 데이터가 부족하기 때문

○ 데이터 분석의 고도화를 위한 환경 조성의 부재 및 빅데이터 분석을 통한 성공사례가 많지 않아 레퍼런스가 부족한 현실

○ 데이터 공급자와 수요자의 탐색비용(searching cost)을 최소화해 시장 참여자를 확대할 수 있는 데이터 거래시장 조성필요

[그림 6-1] 국내 빅데이터 시장 전망

제7장 연구개발성과의 보안등급

보안 등급 분류	보안	일반
		Y
결정 사유	연구책임자 의견	연구기관 자체 검토결과
	해당과제는 민간기업의 SoC공간 내에서 비식별화 실증이 진행되었으며 상용화 시스템을 고도화하여 실증사례를 성과로 하는 실증 과제임	Y

제8장 연구개발과제 수행에 따른 연구실 등의 안전 조치 이행 실적

1. SK텔레콤

가. 고객정보보호센터 규정

○ 본 실험실은 SK텔레콤의 고객정보보호센터(Security Operation Center)이며, 보통 SOC실이라고 칭함. 환경안전관리규정에 의거하여 이화학실험을 수행하지 않는 전기, 설계, 컴퓨터관련 개발실로써 C등급에 해당하는 연구실로 다음의 실험실 환경안전수칙 및 보안유지 서약을 준수함. SoC공간은 다음과 같은 보안관리 규정을 준수하여야 함

- SOC실은 항상 청결하고, 정리정돈상태를 유지하여야 한다.
- SOC근무자 모두 안전수칙 및 안전지침과 보안유지 서약을 준수하여야 한다(※ SKT SOC 시스템 참조)
- SOC근무자는 고객정보취급자로서 고객정보를 취급 시 외부에 유출되지 않도록 보안유지 지침에 준수 한다.
- SOC근무자는 SOC근무신청 절차에 따라 사전에 해당 업무영역 파트장 확인 및 SOC 보안관리자의 근무승인을 득하고, ID Card를 SOC 출입관리시스템에 등록하여 출입한다.
- SOC근무자는 SOC보안관리자 및 출입보안요원의 보안검색 및 통제에 적극 협조하고, 악용 가능한 보안취약점 발견 및 침해/보안사고 발생시 SOC보안 관리자에게 지체없이 보고한다.
- SOC근무자는 본 지침 및 SOC 제반 보안통제규정을 준수하여 SOC 상주 업무를 수행하고, 고객정보를 유출하거나 오남용을 하지 않는다.
- 실험 시작 전에 안전교육과 보안교육을 이수하고, 실험할 내용과 기기의 취급/조작 요령 및 통제사항과 사고발생시 대처 요령을 충분히 숙지한 다음 실험을 실시하여야 한다.
- SOC실에서는 책임자의 지시에 따르고, 무리한 실험은 하여서는 안된다.
- 비상시 행동요령을 숙지하여야 한다(화재 및 응급환자 발생시, 전화기 소화기 및 화재경보기의 위치, 기계기구의 전원 차단스위치 등)
- 소화기의 사용법을 숙지하고, 충약여부를 확인하여야 한다.
- SOC실에서 흡연, 음식섭취, 침식, 놀이 등을 하여서는 안된다.
- 본 과제 수행과 관련하여 연구실 안전 환경 조성에 관한 법률을 충실히 이행
- 본 과제 참여연구원은 SOC실 안전 환경 조성에 관한 법률에 따라 연 2회 교육 및 훈련을 실시
- 비상 시 행동요령 및 SOC실 자체 환경안전점검을 실시하며 사고 피해를 최소화하기 위해 비상 시 행동요령 및 환경안전수칙을 부착하여 이행
- SOC실 안전/보안 담당자를 지정하며 담당자는 매일 SOC실 안전점검 일지를 작성하고 비상 시 출입보안요원(02-6400-7203) 또는 SKT SOC보안 담당자(02-6400-5414) 등에 관련 비상 상황을 신고하도록 함.

※ SKT SOC 시스템의 관련 규정 참조

2. 연세대학교

가. 일상점검사항

당 기관에서 제공하고 있는 안전점검·진단 시스템을 활용하여 일상 점검 및 정기 점검, 정밀안전진단 등을 안전 점검자에 의해 이행됨. 또한 신규 인력에 대한 안전교육 실시를 통해 연구실 내 안전 조치 이행 계획 교육을 통한 수행이 이뤄짐.

(1) 일반사항

일반사항				
연구실험실 정리정돈 및 청결상태	◉	○	○	
연구실험실내 흡연 및 음식물 섭취 여부	◉	○	○	
안전수칙, 안전표지, 개인보호구, 구급약품 등 실험장비(흄후드 등) 관리 상태	◉	○	○	

(2) 전기안전

전기안전				
사용하지 않는 전기기구의 전원투입 상태 확인 및 무분별한 문어발식 콘센트 사용 여부	◉	○	○	
접지형 콘센트를 사용, 전기배선의 절연피복 손상 및 배선정리 상태	◉	○	○	
기기의 외함접지 또는 정전기 장애방지를 위한 접지 실시상태	◉	○	○	
전기 분전반 주변 이물질 적재금지 상태 여부	◉	○	○	

(3) 소방안전

소방안전				
비상구, 피난통로 확보 및 통로상 장애물 적재 여부	◉	○	○	
소화기 표지, 적정소화기 비치 및 정기적인 소화기 점검상태	◉	○	○	
소화전, 소화기 주변 이물질 적재금지 상태 여부	◉	○	○	

(4) 기타(미해당)

○ 소방안전, 가스안전, 화학약품, 기계기구 등의 사항이 있으며 해당 사항은 본 연구실에 해당하는 사항이 아니므로 일상 점검사항에 해당하지 않음

제9장 기 타

1. 빅데이터 비식별화 과제관련 홍보활동

가. 비식별화 조치 가이드라인 및 빅데이터 활용가치에 대한 동영상 컨텐츠 제작 배포

- ○ 챌린지 퍼레이드용 미디어 파사드 동영상 제작
- ○ 빅데이터 비식별화 동영상 제작 배포
- ○ 비식별조치 가이드라인의 이해
- ○ 데이터 마켓플레이스 소개
- ○ 빅데이터의 가치 소개
- ○ 각 솔루션의 특징 및 장점 소개 동영상 제작 배포
- ○ 과천과학관 내 전시용 동영상 제작 전시

나. 과제관련 다양한 미디어 홍보활동 추진

No	기사제목	출처	일자
1	[포럼] 정보보호와 데이터산업의 함수 http://www.dt.co.kr/contents.html?article_no=2017020702102251042001	디지털타임즈	2/6
2	SK텔레콤 “비식별 정보 데이터 산업 이끈다”...개인정보 비식별화 실증 사례 공유 세미나 개최 http://it.chosun.com/news/article.html?no=2833124	IT조선	4/7
3	SK텔레콤, 12일 '빅데이터 개인정보 비식별화' 세미나 개최 http://biz.chosun.com/site/data/html_dir/2017/04/10/2017041001028.html#csidx352ace792ae4692a68501dba891913d	조선비즈	4/10
4	"빅데이터는 21세기 원유"… SKT, 암호화한 비식별 개인정보로 신용 평가 분석 새 지평 열어 http://biz.chosun.com/site/data/html_dir/2017/04/12/2017041202317.html#csidx31b871d7752d8a6aad2796e13c98831	조선비즈	4/12
5	SKT-H보험, 218만명 빅데이터 분석해보니....“이런 고객은 위험” http://www.ddaily.co.kr/news/article.html?no=154876	디지털데일리	4/13
6	통화시간 적을수록 대출연체율도 높다 http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=105&oid=277&aid=0003974921	아시아데일리	4/14
7	[포럼] 4차 산업혁명, 빅데이터 안전성에 달렸다 http://www.dt.co.kr/contents.html?article_no=2017042002102251607001	디지털타임즈	4/19
8	[기고문]비식별 개인 정보 생성 및 거래 현장 적용 실증 연구 http://www.imaso.co.kr/?p=16832	마이크로 소프트웨어	4/20
9	[기고문] 누구나 말하지만 아무도 모르는 4차 산업혁명 http://www.etnews.com/20170420000117	전자신문	4/21

다. 해외 전문가 초청 및 세미나 개최

○ (목적) 빅데이터 개인정보 비식별화를 통한 자료 생성 및 유통 실증과 관련한 세미나를 개최하고, 이에 대한 대내외 홍보를 진행하여 제4차 산업의 핵심인 빅데이터 산업의 활성화를 도모

○ (사업 내용) 빅데이터 개인정보 비식별화 자료 생성 및 유통 관련 세미나 2회 개최 및 빅데이터 관련 개인정보 비식별화를 통한 실증 사업에 대한 홍보

<표 9-1> 추진내용

‘16.10.18~	17.2.10	~17.4.11	‘17. 4.12
개인정보 비식별화 관련 실증 연구	1차 빅데이터 관련 세미나 개최	비식별조치 결합지원 설명회	2차 오픈톡투게더 연계 빅데이터 비식별화 세미나 개최
세미나 기획 및 준비	해외세미나 및 서강대 MLRV행사 연계	인터넷진흥원 연계	미래부, KISTEP 연계

○ (빅데이터 세미나 개최) 4차 산업의 핵심인 빅데이터 산업의 활성화를 도모하고자, 2회 주관 세미나 및 2회의 타기관 연계 세미나 개최

○ 빅데이터 분야 국내외 전문가와 함께 빅데이터 및 비식별화에 대한 지식, 열정, 비전공유의 장 마련

○ 빅데이터 개인정보 비식별화’ 자료 생성 및 유통 실증과 관련하여, 우리나라 빅데이터 산업이 나아가야 할 바람직한 미래상을 전달

○ 빅데이터 개인정보 비식별화 관련 자료 생성 및 유통 실증 등에 대한 내용 홍보를 통해 빅데이터 저변을 확대하고 산업 활성화를 도모

[그림 9-1] 해외 전문가 세미나 개최(2017.2.10.)

[그림 9-2] 빅데이터 비식별화 실증 세미나(2017.4.12.)

라. 챌린지퍼레이드/ 과천과학관 전시/ 대전 창조경제 혁신센터 홍보

대전창조경제센터 스타트업 포럼 비식별 데이터 활용 소개 현장

• 대전 창업포럼 현장에서 비식별 데이터 활용 소개

NUGU를 활용한 빅데이터 비식별화 체험

2. 개인정보 비식별 자료 생성·유통의 현장 적용을 위한 검증 회의

가. 실증 자문위원회

자문 위원명	소속	직위	연혁
김신곤	광운대학교	정교수	국무총리산하 공공데이터 전략위원회 이용 활성화 전문위원장
권규현	한양대학교	조교수	공공데이터 전략위원회 빅데이터 전문위원회 위원
박남훈	안양대학교	정교수	공공데이터전략위원회(KISA) 개인정보 적정성 평가위원
고환경	법무법인 광장	변호사	개인정보보호법령해석자문위원회 위원 미래창조과학부 빅데이터 이용활성화 법제 개선 TF 실무단
권영실	법률사무소 혜율	변호사	한국정보화진흥원 심의위원 한국콘텐츠진흥원 콘텐츠비즈니스 자문위원
전응준	법무법인 유미	변호사	한국인터넷진흥원(KISA) 법률자문 공공데이터전략위원회(국무총리 소속) 실무위원회 위원

(1) 기타 참석위원

자문 위원명	소속	소속부서	직위(업무)
정영수	행정자치부	개인정보보호정책과	행정사무관
원세연	개인정보보호위원회	조사과	서기관
김민철	개인정보보호위원회	조사과	주무관
노명선	한국인터넷진흥원	개인정보비식별조치지원센터	단장
김동현	한국인터넷진흥원	개인정보비식별조치지원센터	연구원
고종오	행정자치부	공공정보정책과	전산사무관 (빅데이터 활성화)

(2) 종합 자문의견('[별첨1]자문의견서' 원본 참고)

○ m-유일성 추가 검증 방식은 민감속성에 기반한 공격을 방어하기 위한 방법으로 적절해 보이며 현행 가이드라인에서도 이와 유사한 방식이 적용되어 있다면 국제적인 기준의 KLT방식 용어 정의에 맞게 가이드라인을 수정할 필요가 있음

○ K와 L의 범위를 다양하게 하더라도 재식별 불가를 증명할 수 없으므로 m-유일성을 고려하고 이를 적용하여 활용성을 증가시킬 것을 권고함

○ 원본 유사도의 선정방식(범주형 데이터의 경우)을 달리하여 활용도를 좀 더 고려해야 할 것으로 보이며 기존 KLT 기법은 재식별 위험이 기술적으로 있음

○ m-유일성 적용시 재식별 '불가능'하다는 서술은 용어적으로 수정이 필요할 것으로 보이며 수치형 민감정보라는 부분은 개인정보보호법상, '개인정보'의 범위에 해당하지 않는 정보로서, 비식별 대상정보가 아니라는 점은 유의할 필요가 있음

○ 대체 임시키를 사용하지 않는 직접 결합 방식은 재식별 가능성을 원천적으로 봉쇄할 수 있을 것으로 판단되며 이와 같은 기술 적용이 가능하다면, 시민단체 등의 문제제기 가능성이 낮아질 것으로 판단됨

○ 직접결합방식이 결합가능성에 의한 개인정보 취급되지 않도록 주의가 필요할 것으로 보이며 실제 데이터를 통함 검증으로 비식별화를 통한 빅데이터의 활용도를 높이는 부분도 검토해 나가야 할 것

나. 실증 회의 사진

[그림 9-3] 개인정보 비식별 자료 생성·유통의 현장 적용 실증 내용 발표

[그림 9-4] 개인정보 비식별 자료 생성·유통 검증을 위한 토의

별 첨 자 료

1. 연세대학교 비식별 비교분석 실증 자문의견서

개인정보 비식별 자료 생성·유통의 현장 적용을 위한 실증

자 문 의 견 서

성 명	전응준	소 속 기 관	유미 법무법인
소 속 부 서	변호사	직 함	

○ 본 안표는 m-유인성이라는 새로운 메트릭스를 제안하고 있는 바, 하나의 레코드에서 식별자는 제외한 모든 속성값에 대하여도 다양성 기준을 적용하라는 취지는 적정·타당하다고 생각됨.

○ 재식별 '본가능' 이라는 개념은 합리적 수준에서 재식별이 매우 어렵다는 의미로 해석되어야 할 것이 타당함. 다만 보유한 '데이블 내에서' 식별이 '본가능' 하다는 표현은 ~~불~~ 가능하다고 생각됨.

2017년 04월 06일

작 성 자 : 전응준 (서명)

개인정보 비식별 자료 생성·유통의 현장 적용을 위한 실증

자 문 의 견 서

성 명	고 한경	소 속 기 관	광장
소 속 부 서	N/A	직 함	변호사

✓ m-유일성 적용시, 재식별 '불가능'하다는 서술은 ~~적절~~ 원칙적으로 수긍이 될만할 것으로 판단됨.
재식별 위험도가 낮아진다는 점은 서술에 적절함.

✓ 수치형 민감정보 또한 넓은 개인정보로서,
'개인정보'의 범위에 해당하지 않는 정보로서,
비식별 대상정보가 아니라는 점은 유의할 필요있음.

✓ 임시 대체키를 활용한 적정성이 필요 없으며,
적법 결합이 가능하다는 점이 기술적으로 가능하다면,
시민단체들의 문제제기의 가능성이 낮아질 것으로
판단됨.

2017년 04월 06일

작 성 자 : 고 한경 (서명)

<u>개인정보 비식별 자료 생성·유통의 현장 적용을 위한 실증</u>

<u>자 문 의 견 서</u>

성 명	권 규현	소 속 기 관	한양대학교
소 속 부 서	기술경영대학원	직 함	~~[illegible]~~ 교수

KLT 방식의 재식별 가능성에 대하여
M-유용성을 통해서 보았음.
M-유용성의 유용성은 충분히 있는 것으로 판단됨.
재식별 가능성은 수학적으로 증명이 가능한 것으로
판단됨.
현재 가이드라인에서 L-다양성을 모두 적용했을
경우와 동일함.
다만 유사도의 산정 방식 (범주형 데이터의 경우)
을 정리하여 유용도를 높여 ~~[illegible]~~ 고려해야 함.

2017년 04월 06일

작 성 자 : 권 규현 (서명)

개인정보 비식별 자료 생성·유통의 현장 적용을 위한 실증

자 문 의 견 서

성 명	정 영 선	소 속 기 관	법무법인 해율
소 속 부 서		직 함	변호사

- 비식별데이터의 안전한 유통을 위해서는 재식별 가능성은 합리적인 수준하에서 불가능하게 만드는 것이 중요한 점검이다. 본 실증 연구는 통해서 제안한 방법은 첫번째로 m-유의성 추가 검증과 두번째로 데이터 결합 방식의 다양화 이다.

- 1종 첫번째 m-유의성 추가 검증 방식은 안전성에 기반한 공격을 방어하기 위한 방법으로 적절해 보인다. 단, 현행 가이드라인에서도 이와 유사한 방식이 적용되어 있다면 국제적인 기준의 KLT 방식 등의 점의에 맞게 가이드라인을 수정하는 것도 필요해 보인다.

- 두번째 직접 결합 방식의 도입에 의한 결합 방식의 다양화 방법은 일시대체키는 생성하지 않는다는 특징에 의해 재식별 가능성이 현저히 낮아진다는 점에서 좋은 해결책이 될것으로 생각한다. 단, 직접 결합 방식이 결합가능성에 의한 개인정보로 취급되지 않도록 주의가 필요해 보인다.

- 결론적으로 m-유의성 방식의 추가검증으로 재식별 가능성이 현저히 낮아지고 직접결합 방식의 도입도 해결책으로 보인다. 단, 실제 Data는 통한 검증으로 비식별화는 통한 빅데이터의 활용도는 높아진 부분도 많이 검토해 나가야 할것으로 보인다.

2017년 04월 06일

작 성 자 : 정 영 선 (서명)

개인정보 비식별 자료 생성·유통의 현장 적용을 위한 실증

자 문 의 견 서

성 명	김 신곤	소 속 기 관	강원대학교
소 속 부 서		직 함	교수

- KLT 방법을 준식별자에게만 적용할 경우 재식별 위험이 있으므로 k-유명성을 본 연구에서 제시된 적용할 경우 재식별 위험성은 거의 없는 것으로 사료됨.

- 김형 결합 방식은 재식별 가능성을 원천적으로 대체 인사이를 사용하지 않는 용어를 사용하는 것으로 판단됨.

2017년 04월 06일

작 성 자 : 김 신곤 (서명)

개인정보 비식별 자료 생성·유통의 현장 적용을 위한 실증

자 문 의 견 서

성 명	박 남 훈	소 속 기 관	안양대학교
소 속 부 서	컴퓨터학과	직 함	교수

m 유일성 추가 검증

m-유일성 : 인간 속성기반 공격 방어용.

기존 k,l,t 기반 비식별 솔루션에서는 재식별위험성이 존재함.

비식별데이터의 유통, 활용성을 위해 가이드라인의 보완을 의견드림

1. m-유일성 : 기존 k,l,t 기법이 재식별 위험이 기술적으로 있음.
k, l의 범위를 다양하게 하더라도 재식별불가를 증명할 수 없음.
따라서, m-유일성을 고려하고 이를 적용하여 활용성을 증가시킬 수 있도록 함.

2. 직접결합방식 : 기존 간접결합방식에 비해 재식별 가능성을 해결할 수 있음.

2017년 04월 06일

작성자 : 박 남 훈 (서명)

주 의

1. 이 최종보고서는 미래창조과학부에서 시행한 미래성장동력 플래그십 사업의 연구보고서입니다.
2. 이 최종보고서 내용을 발표하는 때에는 반드시 미래창조과학부에서 시행한 사업의 연구개발성과임을 밝혀야 합니다.
3. 국가과학기술 기밀 유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 안 됩니다.