"미래를 창조하는 ICT Innovator"

개인정보 추출기 및 익명화 처리 모듈

본 기술은 TXT 문서 및 HWP 문서에서 개인정보를 찾아 표시하고 사용자 선택에 의해 이를 익명화 처리하는 기술임. 본 기술은 SNS 서비스 서버 및 웹 서비스 서버로 확장하면 개인정보 탐지 및 조치가 가능하여 공공 데이터 개방시 혹시 발생할 수 있는 개인정보 노출 우려를 줄일 수 있는 기술임.

인증기술연구실 담당자 조진만

ETRI 한국전자통신연구원
Electronics and Telecommunications Research Institute

목 차

1. 개발기술의 주요내용(1)

기술개념 및 특징

기술개념

- 텍스트 문서나 한글문서(HWP)상에 존재하는 정형 및 비정형 개인정보를 탐지하고 선택에 따라 개인정보를 익명화하여 저장하는 기술.
  - 개인정보 추출기 모듈 : 13종의 정형 및 비정형 개인정보를 자동으로 탐지
  - 개인정보 익명화 처리 모듈 : 추출된 개인정보중 전체 또는 선택한 개인정보를 마스킹 처리후 파일로 저장

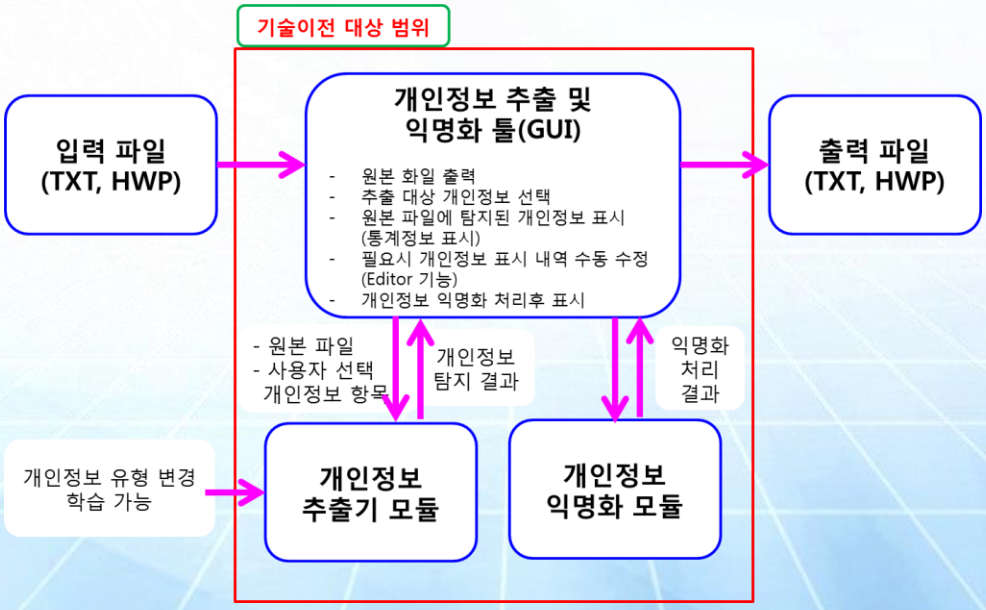
고객/시장의 니즈를 충족시키는 독특한 점(특장점)

- 빅데이터 환경 도래에 따른 정부 3.0의 공공 데이터 개방시 개인정보 유출 우려를 줄일 수 있음
  - 비정형 개인정보의 탐지 : 기존 주민번호, 전화번호, 이메일 등 정형 개인정보의 탐지만으로는 화일내 다양한 개인정보의 유출을 방지할 수 없음
  - 탐지 개인정보의 마스킹 처리 : 데이터의 단순 삭제보다 데이터의 형태를 유지한 마스킹 방식이 공개 파일의 활용성을 높일 수 있음
  - 탐지 대상 개인정보의 변경/조정 및 마스킹 방법의 변경이 가능하여 다양한 환경에서 활용이 가능함

1. 개발기술의 주요내용(2)

기술의 세부내용

- 두개의 핵심 모듈과 1개의 GUI 모듈로 구성.
  - 개인정보 추출기 모듈 : 현재 13종의 개인정보 유형을 화일내에서 탐지하는 모듈
  - 개인정보 익명화 처리 모듈 : 탐지된 개인정보중 사용자 선택에 따라, 전체 또는 특정 유형의 개인정보만 마스킹하는 모듈
  - GUI : Editor 기능 / Viewer 기능 / 파일 저장 기능



기술구성도

1. 개발기술의 주요내용(3)

기술의 세부내용(계속)

- (비정형) **개인정보 추출기 모듈**
  - 현재 13종의 개인정보 유형 탐지
  - 경량 개체명 추출 기능
    - 대형 개체명 사전을 배제, 엔진을 경량화
  - 비정형 개인정보의 정규화 기능
    - 국문 이름 정규화(성이 뒤에 있는 경우 반영)
    - 주소 / 학교명 정규화
  - 자동화된 기계학습 기능
    - 학습 모델 교체시, 새로운 개인정보 유형에 대응 가능
- 탐지 개인정보 **익명화 처리 모듈**
  - 탐지된 개인정보중 사용자 선택 개인정보의 익명화 처리 기능
- **GUI 모듈**
  - TXT 파일 및 HWP 파일 오픈을 통한 화일내 문장 추출 기능
  - 개인정보 탐지 대상 선택 기능
  - 문장 Viewer / 문장내 개인정보 Editing 기능
  - 개인정보 태깅 및 익명화 처리후 파일 저장 기능(TXT, HWP)

탐지대상 13종 개인정보 유형

순번	개인정보 유형	속성/특징
1	사람 이름	
2	주소	광역/시군구/읍면동
3	장소	산/강/섬/관광지 등
4	건물	공항/백화점/호텔/박물관 등 직장명/회사명 포함
5	학교	초/중/고/대/대학원
6	직업	
7	직위	과장, 부장 등
8	주민번호	
9	자동차번호판	
10	전화번호	
11	이메일 주소	
12	나이	
13	날짜	생일, 기념일

1. 개발기술의 주요내용(4)

경쟁기술대비 우수성

경쟁기술/대체기술 현황

- 정형 개인정보 검색 및 마스킹 기술
  - 한글 워드프로세서(HWP) 및 일부 상용 제품에서 화일내 정형 개인정보를 검색하고 마스킹해주는 기술이 존재함
  - 다만, 주민번호, 전화번호 등 정형 데이터에만 적용이 가능하여, 개인정보의 대부분을 차지하는 비정형 개인정보는 탐지가 불가능함
  - 또한 상용 제품으로 학습 등을 통한 새로운 개인정보 유형의 적용이 불가능함.

경쟁기술/대체기술 대비 우수성

경쟁기술	본 기술의 우수성
㈜한글과컴퓨터 / 한글 (HWP)	• 기존 기술에서는 적용이 불가능한 사람 이름, 출생지, 학교, 직업, 직장명 등의 탐지 및 마스킹이 가능함 • 사전에 주어진 항목 이외에, 개인정보 추출기 학습 모델 교체를 통해 새로운 유형의 탐지 및 마스킹이 가능함
㈜이스트소프트 / 닥스키퍼	• 기존 기술에서는 적용이 불가능한 사람 이름, 출생지, 학교, 직업, 직장명 등의 탐지 및 마스킹이 가능함 • 사전에 주어진 항목 이외에, 개인정보 추출기 학습 모델 교체를 통해 새로운 유형의 탐지 및 마스킹이 가능함

1. 개발기술의 주요내용(5)

기술의 완성도

기술개발 완료시기

- 2015년 1월.
  - 비정형 개인정보 추출기 : 2014년 10월 현재 90% 공정
  - 개인정보 익명화 툴 및 GUI : 2014년 10월 현재 100% 완성

기술이전 범위

- 비정형 개인정보 추출기 모듈(원천 코드 및 실행 파일)
- 탐지된 개인정보 익명화 처리 모듈(원천 코드 및 실행 파일)
- 개인정보 추출기 및 익명화 GUI 모듈(원천 코드 및 실행 파일).

1. 개발기술의 주요내용(6)

표준화 및 특허

관련 기술의 표준화 동향

- NBD-PWG
  - 미국 NIST에서 Vender-Independent 표준화 작업을 2013년도부터 시작 (http://bigdatawg.nist.gov)
  - 5개의 Working Group 운영
  - 2014.10. Document 1.0 Release
  - IEEE 및 ISO/IEC JTC1상의 국제 표준화 작업 주도
  - 2015년부터 표준 활용 및 표준 적합성 검증 관련 작업이 예상

보유 특허

출원/ 등록 구분	특허명	출원국 (등록)	출원(등록)번호	출원(등록) 일자
출원 예정	비정형 개인정보 추출 및 익명화 처리 방법	대한민국	-	10월중 출원 예정

2. 기술적용 분야 및 기술의 시장성(1)

기술이 적용되는 제품 및 서비스

기술이 적용되는 제품/서비스

- 정보보안시스템에 개인정보 탐지 모듈로 활용
  - 개인정보 누출/유출 탐지
  - 공공 데이터 개방전 개인정보 포함 여부 확인 및 조치후 공개 처리

- 단품 모듈로 온라인 서버에 적용
  - 개인정보 탐지 뿐만 아니라, 기업명, 상품/제품/서비스의 평판 조회 및 신속한 대응
  - 정치인/연예인 등의 평판 조회 및 평판 관리
  - 여론 조사 및 인기도 조사

- GUI의 데이터 편집기(Editor) 기능 활용
  - 수동으로 특정 단어 또는 유형의 지정 가능
  - 사용자 맞춤형 서비스(특정 단어 입력시 골든벨 작동 등) 제공으로 서비스 활성화

2. 기술적용 분야 및 기술의 시장성(2)

해당 제품/서비스 시장 규모 및 국내외 동향

해당 제품/서비스 시장 규모

- 세계 빅데이터 시장 '17년 약 534억불 전망(연평균 60%대 성장률)
- 안전행정부는 정부 3.0 비전에 따라, '17년 7.7억건의 공공정보 공개 예정

[표 1] 세계 빅데이터 시장 현황 및 전망 (2012~2017) (단위: 백만불)

구분	2012	2013	2014	2015	2016	2017	연평균 성장률
시장규모	6,841	10,200	16,800	32,100	48,000	53,400	60.0

해당 제품/서비스 시장 국내외 동향

- 2012년 국제개인정보보호회의(IDCPPC)는 빅데이터에서 가장 문제가 되는 개인 프로파일링 보호조치를 위한 결의문(Profiling Resolution)을 발의함
- Symantec, McAfee 등 메이저 보안 기업들이 엔드포인트, 스토리지 및 네트워크를 통합한 정보유출방지 솔루션을 중심으로 개인정보 유출 방지를 제공함
- 안랩 등 다수의 보안 업체에서 SNS(트위터, 페이스북) 및 P2P 어플리케이션 등 1,000여개 어플리케이션을 감시/제어하는 '애플리케이션 인텔리전스 서비스' 를 통해 개인정보 유출 차단을 제공함

3. 기대효과

기술도입효과

고객이 본 기술을 통해 얻을 수 있는 경제적 효과

- 공공 데이터 개방시 개인정보 보호 시스템/서비스 활용 참여 가능
  - 정부 3.0 시대에 따른 지속적인 사업 수주 가능성 존재

- 빅데이터상 비정형 개인정보/평판정보 탐지 및 대응 시스템/서비스 가능
  - 한글 특성상 외국계 대형 기업의 진출에 한계가 존재.
  - 적극적 기업이 시장 선제 가능

- 개인정보 탐지 뿐만 아니라, 평판/인기도 조사 등 다양한 영역에 적용 가능
  - 학습 모델 교체 및 유형 변경을 통해, 시대 및 상황에 맞게 엔진 수정 적용