7

개인정보 데이터 접근 비정상행위 탐지 기법을 활용한 개인정보 보호 기법 연구

서울여자대학교 ▌김진형* · 김형종

1. 서 론

최근 들어 인터넷이 급속하게 보급 되면서 이를 이용하는 애플리케이션이 등장하였다. 이러한 애플리케이션은 대용량의 트래픽을 발생 시킨다는 특성이 있다. 최근 발생한 서비스 거부 공격 등의 행위와 같이 대용량의 트래픽이 발생하는 시스템을 임의로 만들어 네트워크 가용성을 떨어뜨리는 공격이 빈번하게 발생하고 있다. 이러한 비정상 행위 탐지 분석을 통해 개인정보 시스템에 접근하는 사용자에 대한 판단을 수행하여 개인정보를 보호할 수 있다. 현재까지 개인정보에 대한 적극적인 보호 노력은 최근 발생한 다양한 개인정보 침해 사고가 갖는 파급력으로 인한 사후 대책의 성격을 띠는 것이었다. 이런 사후적 접근이 갖는 장점은 기존의 발생된 문제들에 대한 해결책 제시가 가능하다는 것이며, 이는 상대적으로 발생하지 않았던 문제들에 대한 답은 아니라는 것이기도 하다.

본 연구에서는 개인정보의 생성, 제공, 사용, 파괴의 각 단계 중 사용의 관점에서 개인정보의 침해를 탐지할 수 있는 모델을 제시하는 것을 목적으로 한다. 개인정보의 정상적인 사용자 모델을 제시하고, 이에 위배되는 상황을 탐지하는 모델의 개발이 그 핵심에 있다. 이러한 모델이 갖는 특성은 개인정보의 침해 여부를 예측할 수 있다는 것으로 사후 약방문 식의 개인정보보호와 접근 방법이 틀리다고 할 수 있다.

본 논문이 제시하는 데이터 모델링 기법은 다양한 웹사이트를 통해서 개인정보를 수집 운영 하는 기관에서 활용가능하다. 본 연구에서는 실제 우리가 사용하는 웹 서비스 내에서 발생할 수 있는 사례를 기준으로 개인정보의 사용 빈도수를 가정에 따라 생성하여 이를 모델링하여 개인정보의 사용빈도 관련 모델을 도출하였다. 이러한 모델의 사용을 통해 개인정보의 일정 기간 동안의 사용빈도를 통해 해당 접근들이 정상적이었는지 아닌지를 탐지할 수 있다.

본 논문의 2장에서는 본 연구에서 접근방법에 대한 관련연구를 제시하고, 일반적인 개인정보의 등급 분류체계 및 사용자분류를 제시한다. 3장에서는 본 연구에서 제시하는 모델링 기법을 제시하고, 일상생활에서 일어나는 사례를 기반으로 모델에서 사용될 데이터의 사용빈도에 대한 측정방법을 제시한다. 4장에서는 3장의 모델을 기반으로 난수생성을 통한 제안 모델의 특성을 분석한다. 5장에서는 결론 및 본 연구 결과의 활용성을 제시한다.

2. 기존 연구

2.1 비정상행위 탐지기법

비정상행위 탐지의 기본원리는 정상행위와 비정상행위는 단정적인 혹은 통계적인 특성의 차이가 있다는 것이다. 이러한 비정상행위에 대한 정의는 특정 데이터 입력, 데이터의 통계적 특성, 이미지의 특정 패턴 등의 다양한 정보를 사용하여 컴퓨터 네트워크, 금융, 이미지처리, 의료 및 산업 등의 다양한 분야에서 활용이 가능 하다. 그림 1은 비정상행위탐지의 분류를 보여주고 있다[15].

침입탐지시스템에서 시스템 및 네트워크의 사용에 대한 비정상적인 행위를 모니터링하기 위한 기법으로 주로 사용되고 있다. 이러한 비정상행위 탐지기법은 신용카드의 오용과 같은 악의적인 범죄행위를 탐지하는데도 사용가능하다. 그 밖의 보험료의 청구, 모바일 단말기를 사용한 오용 및 주식의 내부자 거래 등에 효과적으로 적용할 수 있다. 의료분야의 경우, 특정한 질병 발생을 알리고, 특정 장비에서 발생하는 정보를 기반으로 장비의 오류를 탐지하는 용도로 활용

* 학생회원

† 본 논문은 2009학년도 서울여자대학교 컴퓨터과학연구소 교내학술연구비의 지원을 받았음.

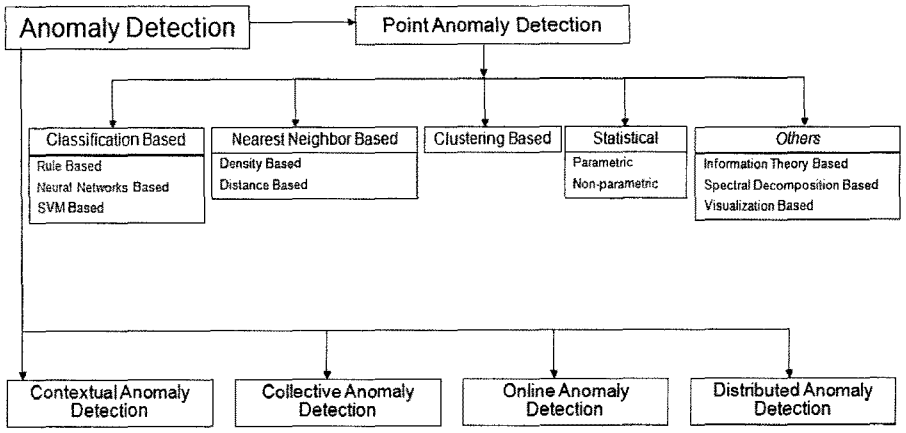

그림 1 비정상행위 탐지 분류도[15]

이 가능하다. 이미지 처리기술에도 위성 이미지에서의 비정상적인 내용을 찾거나 공항 검색환경에서의 비정상적인 상황을 탐지하는 데에 활용이 가능하다.

대표적인 비정상행위 탐지 방법으로 통계적 비정상행위 탐지 방법과 규칙 기반 변화 탐지 방법을 들 수 있다. 통계적 비정상행위 탐지 방법은 시스템에 기록된 명령어를 관찰하는 임계값을 기준으로 하는 탐지 방법과 사용자 행동이 기록된 속성파일 또는 로그파일에 대한 통계적인 탐지 방법을 말한다. 공통 로그 포맷인 CLF(Common Log Format)형식으로 기록되어 있는 웹서버의 로그파일을 분석하여 비정상 행위 지수를 산출하여 적용하는 방법이다. 규칙기반 변화 탐지 방법은 행위기반 도델 중 하나로 과거에 만들어진 감사 자료로부터 얻은 사용자의 정상적인 행위 패턴을 규칙으로 만들어 이에 맞지 않는 경우 침입으로 간주하는 방법을 말한다[16].

2.2 개인정보 데이터 등급 분류

개인정보는 일반적으로 데이터의 민감도에 따라 분류되고 있다. 가장 민감한 등급인 1등급부터 민감하지 않은 5등급까지 5단계로 구분한다. 개인을 식별하거나 개인과 관련된 정보 중에 민감하게 개인과 관련된 정보의 등급이 높은 등급이며, 가장 높은 등급이 P1이고, 가장 낮은 등급이 P5이다. 일반적으로 공개 가능한 데이터의 등급이 P5이며, 본인 이외에는 알 수 없는 데이터는 P1에 해당하게 된다. 표 1은 이러한 데이터들의 등급별 분류를 정리 한 것이다[2,10,11].

표 1 민감도에 의한 데이터 분류

등급	속성
P5 : 사용에 대한 제약이 없는 정보	회원아이디
	회원이름
P4: 개인에 의해 작성된 정보로서, 공개 가능한 정보	회사이름
	부서, 직업
	직장전화번호
	직장주소
	결혼여부
	학교이름
	학위
	자녀수
P3: 개인의 동의하에 공개 및 수집/ 활용 가능한 정보	이메일
	집전화번호
	핸드폰번호
	집주소
	자녀여부
	취미
	혈액형
P2: 필요에 의해 개인의 동의하에 공개 가능한 정보	주민등록번호
	운전면허번호
	여권번호
	거래은행
	계좌번호
P1:절대 공개 되어서는 안되는 정보	패스워드
	카드회사
	카드번호
	유효기간
	수입

2.3 개인정보 사용자 분류

개인정보에 대한 접근 주체가 되는 사용자는 개인정보 사용자, 개인정보 소유자, 개인정보 관리자로 분류 가능하다.

o 정보 사용자(User) : 개인의 정보를 사용하는 사람들을 말한다. 개인정보 데이터에 대해 정보의 소유자에게 요청, 혹은 이미 DB화 되어있는 개인정보 시스템을 관리하는 시스템 관리자에게 개인정보를 요청하여 개인의 정보를 얻음. 개인정보를 활용하여, 본인의 이득을 취하거나, 마케팅,

맞춤정보 제공 등 어떠한 목적으로 사용할 수 있다.

- o 정보 소유자(Owner): 개인정보의 소유자를 의미한다. 타인의 요청에 의해 자신의 정보를 제공하는 정보의 소유자 이다.
- o 정보 관리자(Administrator): 개인정보에 대한 관리책임을 지고 있는 주체를 말한다. 정보에 대한 접근 권한을 대부분 가지고 있으며, 암호화 된 데이터 이외의 정보는 확인 가능함. 접근 데이터의 종류가 많으므로, 높은 수준의 접근 제어 필요하다[12,13].

3. 개인정보 접근 빈도 모델

3.1 로그 데이터 분석 기반 사용자 행위 분석 방법

개인정보를 활용하는 시스템을 운영하는 웹 서버에 접근을 시도하여 개인정보 데이터를 사용하고자 한 사용자의 행위를 분석하기 위해 웹서버의 로그데이터를 활용한다. Web Access Log 데이터를 확인 하여 운영하는 웹서버에 접근 한 사용자의 행위를 분석할 수 있다. Web Access Log를 구성하는 요소들 중 다음 표의 항목을 확인하여 개인정보 활용에 대한 시도를 확인할 수 있다.

이름	Description
Session ID	각 사용자 세션을 위한 Random 승인 ID
Source URL	사용자가 접근을 시도했던 URL 주소
Destination URL	최종적으로 사용자가 접근 했던 URL주소
Stay time	사용자가 Source page에 머물렀던 시간정보

시스템에 접근하는 사용자의 행위를 분석하기 위해 일정 기간 로그 데이터를 수집 하여 그 기간의 Web Access Log를 분석한다면 사용자의 접근 행위에 대한 분석을 수행할 수 있다.

3.2 개인정보 접근빈도 기반 비정상행위 정의 방법

개인정보의 민감도가 높다는 것은 해당 정보가 유출될 경우 관리자 및 소유자에게 미치는 영향이 큰 정보라고 할 수 있다. 이러한 정보에 대해서 관리자는 사용빈도에 대한 지속적인 모니터링을 통해 통계정보를 도출하고, 이를 기반으로 비정상적인 수준의 사용빈도가 발생 시 이를 탐지할 수 있어야 한다.

개인정보에 대한 사용빈도는 개인별 빈도계산과 그룹별 빈도계산을 수행할 수 있다.

- o 개인별 빈도계산 : 각 개인의 정보별로 사용빈도를 계산하여 보유한다. 관리의 대상 단위가 각 개인으로 낮아지며, 각 개인의 정보에 대한 비정상적인 접근을 탐지 및 통보 할 수 있다. 개인정보에 대한 사용빈도 역시도 프라이버시가 될 수 있으므로 이를 수집 관리하는 행위에 대한 별도의 허가가 필요하다.
- o 그룹별 빈도계산 : 특정 웹 사이트의 사용자에 대한 사용빈도를 계산한다거나, 특정 기관의 회원정보에 대한 사용빈도를 계산하는 형태로 진행된다. 이러한 계산 결과를 통해 해당 기관이 보유하고 있는 개인정보가 일정 수준이상으로 사용하는 경우에 대해서 탐지 및 통보 할 수 있다. 그룹의 범위를 국가로 확대할 경우, 국가의 개인정보 사용빈도의 급작스러운 증가를 기반으로 비정상 상태를 탐지할 수도 있다.

수식 (1)은 특정 개인정보에 대한 사용빈도 계산식이다.

$$Freq\_Usage(i,j) = Num\_Usage(i,j)/T \quad (1)$$

단, Freq_Usage(i,j) : i개인정보의 j관측 시간별 T 시간당 사용빈도
Num_Usage(i,j) : i개인정보의 j관측시간 내에서의 사용횟수
T : 사용주기를 계산하기 위한 단위시간

수식 (1)에서의 j관측시간은 관리자가 중요시 하는 관측시간으로(Observation Interval)로 이 시간에 관측된 내용을 가지고 통계를 도출한다. 단위시간 T는 관리자가 정의한 단위시간으로 사용빈도를 계산하기위한 값이다.

이렇게 계산된 사용빈도는 지속적으로 수집하여 다음 식을 활용한 정규분포를 생성할 수 있다. 이렇게 만들어진 정규분포를 활용하여 개인정보에 대한 정상적인 정보 사용빈도에 모델을 수식 (2)와 같이 얻을 수 있다.

$$p(x_i) = \frac{1}{SD_i^{Freq}\sqrt{2\pi}} exp(-\frac{(x_i - Mean_i^{Freq})^2}{2SD_i^{Freq}}) \quad (2)$$

$p(x_i)$: x_i에 대한 확률 밀도 함수
x_i : 최근 관측된 정보 i에 대한 사용빈도
SD_i^{Freq} : 정보 i에 대한 사용빈도의 표준편차
$Mean_i^{Freq}$: 정보 i에 대한 사용빈도의 평균

수식 (1)과 (2)에서 제시된 사용빈도에 대한 수식 모델은 모든 데이터에 대해 관리자가 지정한 관측시간 및 단위시간에 근거해서 지속적으로 계산되어 정규분

포를 구성하는 평균과 표준편차에 반영된다. 이를 통해 수식 (2)의 확률밀도 함수를 도출할 수 있고, 일반적인 데이터에 대한 사용 접근빈도를 계산할 수 있다. 확률밀도함수의 출력으로 얻어지는 값은 최근 관측된 사용빈도의 확률적 위치를 알려주고, 관리자는 이 상황에 대한 정상/비정상을 판단하게 된다.

그림 2는 도출된 정규분포와 확률밀도 함수를 사용하여 비정상상태를 탐지하는 예를 보여주고 있다. 본 개인정보 사용빈도와 관련하여, 일정 빈도 이상으로 사용된 경우에 대해서 비정상 상태라고 볼 수 있으며, 이를 전체의 몇 퍼센트에 해당하는 지에 대한 수치를 통해 나타내게 된다.

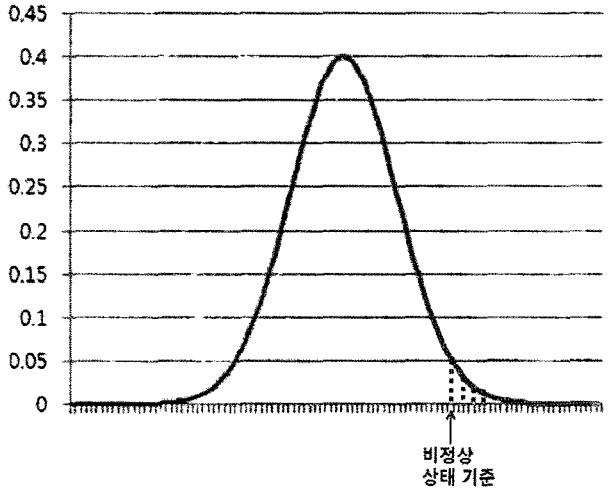

그림 2 정규분포와 확률밀도함수를 사용한 비정상 상태 탐지

그림 3은 이러한 개념을 활용한 시스템의 구성 예이다. 데이터에 대한 접속 현황은 모니터링 됨과 동시에 새로운 통계모델을 도출하기위한 데이터로 활용된다. 새로운 접속현황은 비정상행위 판정모듈에 전달되고, 통계모델의 내용을 기반으로 하여 비정상 상태여부를 탐지한다. 비정상행위판정 모듈이 가져야 하는 중요정보로 각 데이터별 비정상행위결정 임계치를 가져야한다.

웹사이트의 개인정보에 대한 접근빈도의 모니터링을 통한 비정상행위 탐지를 수행하기 위해서는 각 개인정보에 대한 접근빈도를 산출하여 통계적 모델을 만들고, 이를 지속적으로 유지하는 것이 중요하다. 3.3절은 다양한 웹사이트의 각 개인정보에 대한 빈도 산출 방법을 제시하고 있다.

3.3 데이터 별 접근빈도 산정을 위한 웹 사이트 환경 정의

데이터 별로 사용되는 빈도수에 대한 산정을 위해 본 논문에서는 Web에서 사용된 데이터를 기준으로 1주일 동안 6개의 사이트에서 발생하는 8가지의 사례를 고려하였고, 각 데이터가 사용되는 횟수의 평균값을 산정하였다. 예를 들어 구매를 대행하는 사이트에서 일어날 수 있는 사례는 회원가입, 로그인, 결재 등이 있다. 회원 가입부터 로그인 횟수 등을 일주일을 기준으로 상대적으로 산정하여 데이터별 사용회수를 산정하였다. 각각의 사례발생은 본 연구를 위하여 최

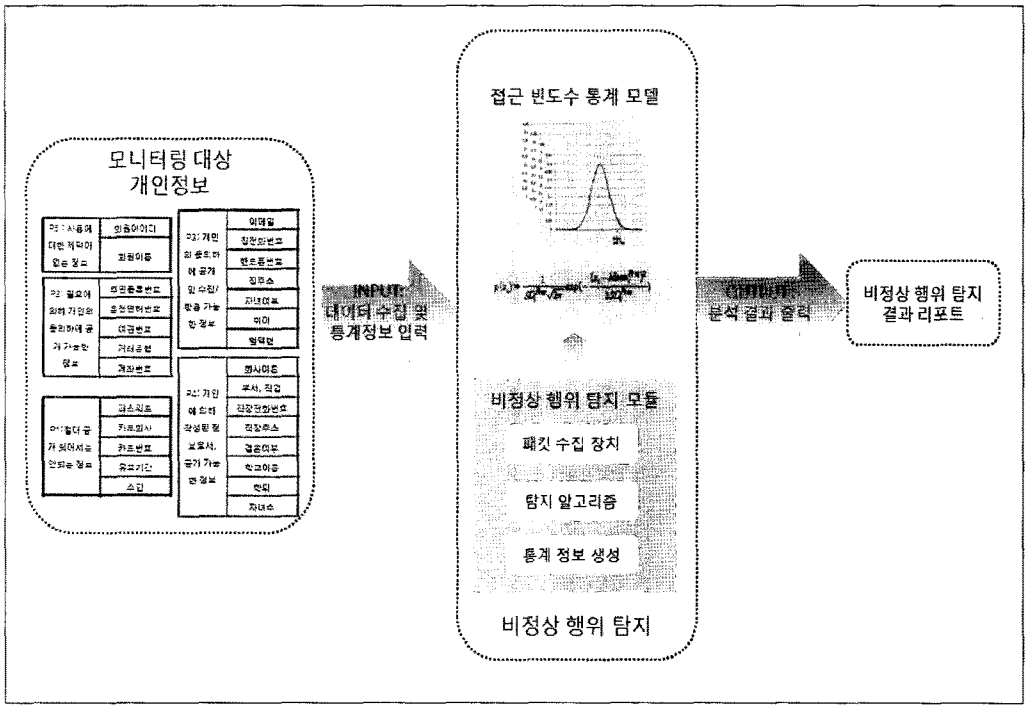

그림 3 비정상행위 탐지 시스템 내부구성

사이트1(S1): 개인 간 구매전문 사이트
사이트2(S2): 검색엔진, 지식검색 사이트
사이트3(S3): 검색엔진 사이트
사이트4(S4): 클럽, 블로그 사이트
사이트5(S5): 카페, 클럽활동
사이트6(S6): 개인 간 소호몰 구매전문 사이트

그림 4 개인정보가 활용 되는 대표적인 웹 사이트

사례 1(C1): 회원가입_웹사이트
사례 2(C2): 로그인
사례 3(C3): 비밀번호 분실 재발급
사례 4(C4): 구매/결재정보(회원구매)
사례 5(C5): 구매/결재정보(비회원구매)
사례 6(C6): 직통연락정보조회
사례 7:(C7) 개인성향 부가정보 정보조회
사례 8(C8): 취미-운동/음악 정보조회

그림 5 사용 되는 사이트와 사이트 내에서 발생 가능한 사례 분류

대한 객관성을 갖도록 가정을 하였다. 예를 들면, 회원 가입 후 로그인을 하고 결재를 할 수 있는 시스템에서는 회원가입이라는 것은 1회만 발생하며, 로그인은 사이트에 접속할 때 회원에게 제공되는 서비스 사용 시 수시로 필요한 과정이다. 또한, 결재를 하는 경우는 로그인 횟수 보다는 상대적으로 낮은 빈도가 발생한다는 것을 예측할 수 있다. 그림 4는 데이터접근 빈도의 산정을 위해 고려된 여섯 가지 유형의 사이트를 보여주고 있으며, 그림 5는 그림 4의 각 사이트별로 발생 할 수 있는 여덟 가지의 사용 사례를 보여주고 있다[9,11].

표 3은 위의 그림 5에 나와 있는 웹 사이트 내에서 발생 가능한 사례들에 대해 각 사례 발생 시 사용된 데이터의 종류를 정리 한 것이다.

표 3은 실제로 분류되어 있는 각 사례 발생시, 사용되는 데이터를 정리 한 표로 각각의 사이트에서 접근하고 활용하는 데이터가 다를 수 있지만, 전반적으로 사용되는 데이터를 표로 묶어 놓았다. 위 8개의 사례는 사용자의 생활패턴에 따라 빈도가 다를 수 있으며, 8개의 사례의 여러 부분에 속해 있는 정보의 경우 그 접근 빈도가 다른 것에 비해 커질 것이라는 것을 예측 할 수 있다. 또한 경우에 따라서는 일부 사례들은 사례 한번을 실행하기 위해 나열된 개인정보를 2회 이상 접근하는 경우도 발생할 수 있다.

한 가지 더 고려할 수 있는 것은 각 사례별 발생 빈도에 대한 순위이다. 각 사례들은 일반적인 사용자들의 사용 성향을 고려할 때, 사례별로 그림 6과 같은 순위를 고려할 수 있다. 물론 이러한 비교 결과는 실제 데이터를 통해서 얻어야 하지만, 이러한 특성을 갖는다는 것을 직관적으로 알 수 있다. 그림 6의 내용은 본 연구의 모델 제시 및 검증과정에서 가상의 데이터의 특성을 나타내는 용도로 활용되었다.

3.4 사례별 발생 빈도 순위

3.3에서 정리된 사이트의 유형별로 나타날 수 있는 사례를 정리하면 아래와 같다. 각 사이트 별 사례는 사이트가 갖고 있는 기능 및 사용자에게 제공하는 서비스에 따라서 다르게 나타나게 된다. 예를 들면, 모든 사이트들이 회원가입(C_1)의 기능을 갖지만 구매기능(C_4, C_5)의 경우는 모든 사이트가 갖지 못한다.

표 3 분류된 사례 내에서 사용되는 데이터의 종류

회원가입	이름,주민등록번호,주소,핸드폰번호,전화번호,메일주소,ID,취미,회사정보,전공,학력,분야,자녀수,결혼여부
로그인	ID, PWD 메일주소
비밀번호 분실재발급	주민등록번호, 비밀번호 질의응답, 이름, ID, 메일주소
구매(회원가입-로그인후)	핸드폰번호,메일주소,카드정보,은행정보,이름,주소,ID
구매(비회원구매)	주민등록번호.이름.주소.핸드폰번호.메일주소.카드정보.은행정보
직통연락정보	전화번호,이름,메일주소,나이,성별,핸드폰번호
개인성향/부가정보활용	자녀수,결혼여부,이름,결혼기념일,직업,첫출산 나이, 자녀의 생년월일, 성별, 자녀의 성별
취미정보(예: 운동/음악)	관심분야,이름,직업,관심종목,전화번호,메일주소

회원가입 〈 비회원구매 ≤ 취미정보 ≤ 비밀번호 질의응답 〈 부가정보 활용 〈 구매(회원) 〈 Direct Contact 〈 로그인

그림 6 사례별 발생빈도 순위

S_1 = {C_1, C_2, C_4, C_6, C_7, C_8}
S_2 = {C_1, C_3, C_4, C_7, C_8}
S_3 = {C_1, C_4, C_7, C_8}
S_4 = {C_1, C_3, C_4, C_5, C_7, C_8}
S_5 = {C_1, C_2, C_3, C_5, C_7, C_8}
S_6 = {C_1, C_2, C_4, C_5, C_6, C_7, C_8}

단, C_n = 발생 가능한 케이스 S_n = 사이트

본 논문에서는 위와 같은 사이트별 사례특성을 기반으로 각 사이트에 나타나는 사례들에 대한 빈도계산을 예를 들어서 수행해보고자 한다. 표 3은 1인이 1주일 동안 6개의 사이트에서 발생시킬 수 있는 사례의 개수를 이해를 돕기 위해서 예시한 것이다. 사이트의 특성에 따라 특정사례들은 나타나지 않을 수 있으며, C_8에서처럼 매우 많이 나타날 수도 있다.

이를 사용자 관점에서 정리하면, 표 4와 같이 정리할 수 있다. 즉, S_1, S_4의 사용자의 경우 2개 사이트의 사용 회수를 더하여, 개인의 정보가 노출되는 회수의 근거가 되는 정보를 산출 할 수 있다.

3.3절의 표 3에서 명시한 바와 같이 각 사례들은 발생할 때마다 참조되는 정보들이 결정되어 있다. 따라서 해당 사례의 발생은 사례와 연결되어 있는 개인정보에 대한 접근이 발생한 것을 의미한다. 이를 표

표 4 6개 사이트에 대한 1 주간 1인 생활 패턴에서 발생 가능한 사례의 수

사례 / 사이트	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
S_1	1	1	2	0	0	15	10	24
S_2	1	0	2	7	0	0	10	30
S_3	1	0	3	0	0	0	15	30
S_4	1	0	2	9	8	0	20	30
S_5	1	3	3	0	5	0	14	25
S_6	1	3	2	0	10	15	20	25
합계	6	7	14	16	23	30	89	164

표 5 일부 사이트를 이용하는 특정 사용자 기준 사례 발생 수

사례 / 사용자	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8
U_1 (S_1, S_4)	2	1	4	15	12	15	30	54
U_2 (S_1, S_2, S_5)	3	4	7	5	5	15	34	79
U_3 (S_1, S_3, S_4)	3	1	7	15	12	15	45	84
U_4 (S_5, S_6)	2	6	5	0	15	15	34	50

시하기 위해서는 위의 표와 같은 회수 정보와 각 사례가 발생시키는 정보접근 회수 정보가 요구된다.

표 2를 통해 볼 때 주민등록번호의 경우 3가지 사례(C_1, C_3, C_5)에서 사용되고 있다. 따라서 U_1의 경우에 대한 주민번호 접근 회수를 계산하면 아래와 같은 수식 (3)을 통해 계산이 가능하다.

$$AccessNum_{U_1}^{SSN} = OC_{U_1}^{C_1} + OC_{U_1}^{C_3} + OC_{U_1}^{C_5} \quad (3)$$

$AccessNum_{U_1}^{SSN}$: U_1의 주민번호(SSN) 접근 횟수
$OC_{U_i}^{C_j}$: U_i 사용자의 C_j 사례 경우 발생 횟수

4. 모델의 활용 예시

본 장에서는 데이터 별 평균값 예시를 기준으로 데이터별 사용분포를 알아본다. 표 5는 앞에서 분류한 6개의 사이트에 대해 8개의 사례 발생 시 사용된 데이터를 기준으로 횟수를 측정하여 각 데이터 별로 평균을 알아보았다. 해당 평균정보는 가정에 의해서 도출된 것이지만, 본연구의 접근 방법에 대한 검증을 위한 활용성 측면에서는 크게 무리를 주지 않는다. 특히, 수치의 도출을 위해서 사용된 가정을 뒷받침해 주는 정보로서 3.3절에 제시된 사례의 빈도 차이가 있다. 일반적으로 웹사이트내 사용자 행위의 빈도는 아래와 같은 값의 차이가 존재한다고 보는 것이다. 이러한 행위 빈도의 비교 정보는 우리가 통상적으로 생각할 수 있는 수준의 내용으로 생각되며, 이러한 정보를 기반으로 생성된 본 연구의 데이터 역시 나름 데로 연구내용의 검증을 위해 논리성을 제공한다고 사려 된다.

각 6개의 사이트에 데이터의 평균값은 아래 표 6과 같으며, 각 사이트에서 사용되는 데이터를 기준으로 데이터별 사용 빈도수와 평균값을 기준으로 정규분포 그래프를 그려볼 수 있다. 또한 정규분포 정보를 활용하여, 난수생성기를 활용하여 표 7과 같이 15종의 샘플 데이터를 생성해 봤다.

5. 결 론

본 논문은 웹사이트에 대한 개인정보 접근 회수를 기반으로 비정상 행위 탐지를 위해서 활용할 수 있는 모델을 제시하였다. 특히, 각 데이터들에 대한 접근 분포 모델을 개발하기 위해 웹사이트의 활용 사례를 8가지로 구분하고 각 사례와 연결된 주요 데이터를 추출하였고, 각 데이터별로 정규분포 모델을 개발하기 위한 방법을 제시하였다.

본 논문에서 제안하는 모델을 개인정보 관리 시스

표 6 데이터별 수행 횟수를 기준으로 한 데이터별 분포도

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
이름	2	4	6.50	38.90	44.30	49.70	55.10	60.50	55.10	49.70	44.30	38.90	6.50	4	2
주민번호	1	2	3.67	9.07	9.97	10.87	11.77	12.67	11.77	10.87	9.97	9.07	3.67	2	1
주소	3	5	9.82	15.43	16.36	17.30	18.23	19.17	18.23	17.30	16.36	15.43	9.82	5	3
핸드폰번호	2	5	8.00	17.00	18.50	20.00	21.50	23.00	21.50	20.00	18.50	17.00	8.00	5	2
전화번호	1	2	2.67	8.67	9.67	10.67	11.67	12.67	11.67	10.67	9.67	8.67	2.67	2	1
메일주소	0.3	0.7	1.48	34.69	40.23	45.76	51.30	56.83	51.30	45.76	40.23	34.69	1.48	0.7	0.3
ID	12	15	20.65	39.46	42.60	45.73	48.87	52.00	48.87	45.73	42.60	39.46	20.65	15	12
PWD	4	7	10.85	26.84	29.51	32.17	34.84	37.50	34.84	32.17	29.51	26.84	10.85	7	4
비밀번호 질의응답	0.5	0.6	0.98	2.99	3.33	3.66	4.00	4.33	4.00	3.66	3.33	2.99	0.98	0.6	0.5
카드정보	2	3	3.98	10.19	11.23	12.26	13.30	14.33	13.30	12.26	11.23	10.19	3.98	3	2
은행정보	1	1.5	2.18	5.57	6.14	6.70	7.27	7.83	7.27	6.70	6.14	5.57	2.18	1.5	1
기타	1	4	7.00	39.07	46.22	53.37	60.52	67.67	60.52	53.37	46.22	39.07	7.00	4	1

표 7 6개 사이트 별 데이터 사용 빈도수

사이트 / 개인정보 데이터	S1	S2	S3	S4	S5	S6
이름	9.5	16.5	20.5	24.5	57.5	13.5
주민등록번호	1.67	2.67	2.67	3.67	5.33	5.33
주소	1.17	1.17	18.17	11.83	7.83	0.83
핸드폰번호	5	5	12	11	7	4
전화번호	7.67	8.67	8.67	13.33	8.33	3.33
메일주소	11.83	18.83	26.83	30.83	83.17	5.17
ID	1	2	17	7	7	6
PWD	2	1	2	2	1	0
비밀번호힌트	0.33	4.33	4.33	5.67	2.67	0.67
카드정보	1.33	3.33	5.33	7.67	1.67	0.67
은행정보	1.83	1.83	3.17	0.17	0.17	0.17
기타	26.67	28.67	35.67	57.67	143.33	5.33

템에 적용한다면, 개인정보에 대한 안전한 관리 및 비정상적 접근에 대한 탐지가 일정부분 실행될 수 있을 것이다.

향후 본 연구결과에서 제시된 개인정보의 접근에 대한 비정상행위 탐지 시스템의 개발이 진행될 것이다. 이러한 시스템의 특성이 탐지 과정의 자동화로 인한 미탐과 오탐이다. 이러한 오류를 보완하기 위한 모델의 조율이 진행될 것이다.

6. Acknowledgement

본 논문은 2009년도 정보보호학회 하계 학술대회 발표한 논문에서 확장된 연구 결과를 정리 한 것임.

참고문헌

[1] 강성철, "개인정보보호 실태와 정책 방향", 한국인터넷정보학회 학회지, pp54-58, 2000.12

[2] 강용석, "개인정보의 정의 및 보호 트랜드", 인포섹 IT Solution, 2005년 9월호 칼럼

[3] 권건보, "개인정보보호와 자기정보통제권", 경인문화사, pp17-20, 2005.12

[4] 김연수, "개인정보보호", (주)사이버출판사, pp31-45, 2001.7

[5] 김우철, 김재주외 8인, "통계학 개론", 영지문화사 (제4개정판), pp120-142, 2002.1

[6] 나석현, 박석, "사용목적 분류를 통한 프라이버시 보호를 위한 접근제어 모델", 한국정보보호학회 논문지 제 16권 3호, pp39-52, 2006.6

[7] 변재옥, "정보화 사회의 프라이버시와 표현의 자유", 커뮤니케이션북스, pp38-73, 1999.6

[8] 송상헌, 이종후, 류재철, "전자상거래 보안", 전자공학회지 제 28권 6호, pp55-64, 2001.6

[9] 송유진, 이동혁, "개인정보 라이프사이클에 따른 프라이버시 보호 프레임 워크", 한국정보보호학회 논문지 제 16권 4호, pp77-86, 2006.8

[10] "안전한 전자정부를 구현하기 위한 개인정보보호 및 정보보안대책", 전자정부특별위원회, 2002

[11] 한국정보보호진흥원, "2002년 개인정보보호백서", 한국정보보호진흥원, 2002

[12] Biswajit Panja, Sanjay Kumar, Bharat Bhargava, "A role-based access in a hierarchical sensor network architexture to provide multilevel security", Computer Conmminications 31, pp793-806, 2008

[13] Dieter Gollmann, "Conputer Security",John Wiley & Sons, 2005

[14] G. Pernul et al., "Modeling data secrecy and integrity", Data & Knowledgr Engineering 26, pp

291-308, 1998

[15] Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Outlier Detection - A Survey", Technical Report TR07-17, University of Minnesota

[16] 김효남, "인터넷 환경에서의 비정상행위 공격탐지를 위한 위협관리 시스템", 한국 컴퓨터정보학회 논문지 제11권 5호, p157-p164, 2006.11

[17] Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru, "Identifying Web Spam with User Behavior Analysis", AIRWeb'08, p9-p16, 2008.4

김 진 형

2006 서울여자대학교 정보보호공학과 졸업
2008 서울여자대학교 대학원 컴퓨터학과 석사
2008~현재 서울여자대학교 컴퓨터학과 박사과정
관심분야 : 정보보호, 개인정보보호, 디지털 포렌식
E-mail : jinny@swu.ac.kr

김 형 종

1996 성균관대학교 정보 공학과(공학사)
1998 성균관대학교 정보 공학과(공학석사)
2001 성균관대학교 전기전자 및 컴퓨터공학과 (공학박사)
2001~2007 한국정보보호진흥원 수석연구원
2004~2006 Carnegie Mellon University, USA Visiting Researcher
2007~현재 서울여자대학교 컴퓨터학부 조교수
관심분야 : 취약점 분석 및 모델링, 이산사건 시뮬레이션 방법론, 인터넷 전화 보안
E-mail : hkim@swu.ac.kr