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

The Journal of Society for e-Business Studies
Vol.21, No.1, February 2016, pp.1-14
http://dx.doi.org/10.7838/jsebs.2016.21.1.001

웹사이트의 구조를 고려한 개인정보 노출 위험도 계산 기법

A Method for Calculating Exposure Risks of Privacy Information based on Website Structures

이수경(Sue Kyoung Lee)*, 손진식(Jin Sik Son)**, 김관호(Kwanho Kim)***

초 록

본 연구에서는 개인정보가 웹사이트에 노출될 시 위험 정도를 수치화할 수 있는 웹사이트 구조기반의 개인정보 노출 위험도 모델을 정의하기 위해 아래와 같은 두 가지 측면을 고려한다. 첫 번째는 개인정보가 노출되었을 경우 얼마나 민감한 정보인가에 따라 위험수준을 정의한다. 두 번째는 개인정보의 실제 노출 가능성을 측정하기 위해 웹페이지의 예상 방문 확률을 계산하여 어느 웹페이지에 노출된 개인정보가 더 위험한지 판별한다. 이를 바탕으로 대학교, 은행, 중앙 행정 기관, 시・도 교육청 4개의 분류를 선정하여 웹사이트 위험도를 측정하였다. 실험 결과, 은행은 다른 분류에 비해 상대적으로 잘 관리되고 있었으며 시・도 교육청, 중앙 행정 기관, 대학교의 경우 웹사이트 위험도가 높게 측정되었다. 마지막으로, 본 연구는 개인정보 노출 문제의 완화를 위한 우선순위 기반 대처방안 수립에 도움을 줄 것으로 기대한다.

ABSTRACT

This research proposes a method that aims to evaluate the risk levels of websites based on exposure risks of privacy information. The proposed method considers two aspects as follows. First, we define the risk levels of each privacy information according to its own inherent risk. Second, we calculate the visiting probability of a webpage to measure the expected of the actual exposure of privacy information on that webpage. In this research, we implemented an system to prove that automatically collects websites and calculates their risk levels. For the experiments, we used a real world dataset consisting of a total of websites for 4 categories such as university, bank, central government agency, and education. The experiment results show that the websites in the bank category are relatively well managed, while the others are needed to cope with the exposure of privacy information. Finally, the proposed method in this research is expected to be further utilized in establishing a priority-based approach to alleviate of the privacy information exposure problems.

키워드 : 개인정보, 노출위험, 웹콘텐츠 분석, 웹링크 분석, 인터넷 웹사이트
Privacy Information, Exposure Risk, Web Contents Analysis, Web Link Analysis, Internet Websites

이 논문은 인천대학교 2014년도 자체연구비지원에 의하여 연구되었음.

* Department of Industrial and Management Engineering, Incheon National University(dltnrud1212@naver.com)
** Integrated Safety Assessment Department, Kepco Engineering & Construction Company, INC. (realrice@kepco-enc.com)
*** Corresponding Author, Department of Industrial and Management Engineering, Incheon National University (khokim@inu.ac.kr)

Received: 2015-11-10, Review completed: 2015-12-19, Accepted: 2016-02-01

1. 서 론

오늘날 웹사이트를 통한 정보 공유가 가속화되면서 개개인의 신상을 파악할 수 있는 개인정보의 노출 문제가 야기되고 있다. 예를 들면, 웹사이트상의 불특정 다수에게 개인정보가 노출될 경우 전화번호나 주민등록번호 도용 등의 타인의 정보 훼손, 침해, 도용 사례가 발생할 수 있다. 특히 의료 및 금융 분야는 매우 민감한 정보를 다루기 때문에 신분도용이나 신용 사기와 같은 위험에 노출되어 금전적인 손해를 입을 수 있다.

또한, 웹사이트 개인정보 노출 위험도를 웹페이지 단위로 분석하는 것은 웹사이트 관리 측면에서도 매우 중요하다. 분석된 웹페이지별 노출 위험도를 바탕으로 위험도가 높은 웹페이지의 우선적 관리를 통해서 웹사이트 위험도를 효율적으로 낮출 수 있는 지표로 활용될 수 있다.

본 연구에서의 특정 웹사이트의 노출 위험도는 해당 웹사이트 사용자들이 방문하여 웹사이트 구조에 따라 웹페이지들을 방문할 때 기대되는 개인정보의 총 노출 횟수로 정의한다. 이때, 웹사이트가 가지는 구조적 특수성으로 인해 동일한 개인정보라도 노출된 웹페이지의 구조적 위치에 따라 개인정보 노출의 심각도가 달라진다. 웹페이지 내에는 다른 웹페이지를 가리키는 링크가 삽입되어 있어, 링크가 가리키고 있는 다른 웹페이지와 자신의 웹페이지 사이의 연관성을 갖게 되므로 특정 웹페이지의 링크 참조가 증가하면 해당 웹페이지가 사용자들에게 노출될 가능성 또한 증가한다.

따라서 구조적으로 방문확률이 높은 웹페이지일수록 포함된 개인정보의 기대 노출 횟수가 증가하게 된다. 또한, 웹사이트 구조상 많은 경우 80%의 방문 트래픽은 전체 웹페이지의 20%에서 발생한다[20]. 이는 동일한 개수의 개인정보를 포함하는 웹사이트도 개인정보를 포함하는 웹페이지들의 구조적 특성에 따라 기대되는 개인정보 노출횟수는 크게 달라질 수 있음을 의미한다.

따라서 본 논문에서는 민감도에 따른 개인정보의 위험수준과 개인정보가 노출된 웹페이지의 예상 방문 확률을 동시에 고려한 개인정보 노출 위험도 계산 모델을 제안한다. 구체적으로 개인정보의 위험수준 정의는 개인정보 노출 시 피해 정도를 의미하는 민감도에 따라 위험수준을 정의한 기존 연구를 토대로 확장한다[9, 19]. 그리고 각 웹페이지의 예상 방문 확률 측정은 특정 웹사이트 내 웹페이지 간의 상호 연결된 링크 정보로부터 웹페이지의 중요도를 평가하는 HITS 알고리즘을 활용하여 제안한다[14].

제안된 개인정보 노출 위험도 기법의 실적용을 위해 대학교, 은행, 중앙 행정 기관, 시·도 교육청 4개 분류를 선정하였다. 웹 트래픽 전문 조사기관에서 제공하는 분류별 웹사이트들을 선정하고[22], 실험 결과를 토대로 분류별 웹사이트의 개인정보 관리 상태 및 계산 모델에 대한 타당성을 분석하였다.

본 논문의 구성은 제 2장에서 기존 연구의 고찰 및 한계점을 분석하고, 제 3장에서 개인정보의 위험수준과 웹페이지의 예상 방문 확률을 고려한 웹사이트 개인정보 노출 위험도 계산 모델을 제시하며, 제 4장에서는 제안 기법을 적용하여 구현한 시스템의 실험 결과를 분석한 뒤, 제 5장에서 결론을 맺는다.

2. 관련 연구

기존의 개인정보 노출 위험도의 분석연구는 <Table 1>에서와 같이 크게 두 분류로 구분될 수 있다.

개인정보 유출 시 자체의 고유 위험수준을 고려한 Kim[9], Lee and Young[15], MOSPA-KISA[19]은 개인정보 민감도에 따라 개인정보에 위험수준을 부여하며, 다수의 개인정보 항목이 조합되어 탐지될 수 있다는 점을 고려한 등급을 제시했다.

또한, Cho and Jun[4], Park and Lim[21]는 개인정보 유출 위험성 감지를 위해 각각의 개인정보 항목(주민번호, 이메일주소, 핸드폰번호 등)이 갖는 정보 민감도나 중요도에 따라 개인정보의 유출 징후에 대한 임계값을 설정했다. Choi et al.[5]는 공개된 데이터에 포함된 특정 개인의 개인정보를 분석하여 노출된 개인정보의 범위와 민감 정도에 따른 위험도를 평가하는 기술을 제안했다. 그러나 개인정보 자체의 위험수준에 관련된 부분만을 고려하고 있으며, 개인정보의 실제 노출될 가능성을 따로 고려하지 않았다.

Lee et al.[16], Cheon et al.[3], Choi et al.[6], Kim et al.[13]는 개인정보의 자체적인 위험수준뿐만 아니라 노출될 가능성도 고려하였다. 나아가서, Lee et al.[16]는 웹사이트 게시물의 특성에 따른 노출 수준과 각 개인정보의 자산 가치를 고려한 개인정보 노출 위험도를 산출했다. 하지만, 게시물과 같은 서비스에 한정되어 있어 웹사이트에 적용 시 어려움이 있으며, 서비스 이용자의 입장에서만 게시물의 접근허가 특성에 따른 위험정도를 고려하였다는 한계점이 있다.

Cheon et al.[3]는 SNS(Social Network Services)에서의 개인정보 유출위험도를 측정하기 위해서 개인을 식별할 수 있는 정도와 개인정보의 악용 정도에 따라 위험정도가 부여된 개인정보 자산 가치와 개인정보 유출가능영역을 함께 고려했다. 이 연구는 각 개인의 중요도가 동일함을 가정하고 거리에 따른 가중치를 설정하지만, 웹사이트의 경우 각 웹페이지는 상이한 방문확률을 지니므로 중요도를 달리할 필요가 있다.

Choi et al.[6]의 연구에서 위협수준은 해당 웹사이트의 노출 페이지 건수, 해당 웹사이트의 신뢰도, 위협지수를 고려하여 산정되며, Kim et al.[13]의 개인정보 위험도 산출은 프라이버시 민감도와 개인 식별성을 기준으로 한 절대적인 노출 위험도를 고려하고, 또한 공격자의 정보수집능력 및 관심정보를 기준으로 하는

〈Table 1〉 Analysis of Existing Research from the Research's Perspective

The Research's Perspective	Descriptions	Reference
Risk level of privacy	In case of privacy exposure, the risk is measured considering the inherent risk level of privacy.	Cho and Jun[4]
		Choi et al.[5]
		Kim[9]
		Lee and Young[15]
		MOSPA-KISA[19]
		Park and Lim[21]
Exposure probability of privacy	The risk is measured considering not only the inherent risk level of privacy but also the exposure probability of privacy.	Cheon et al.[3]
		Choi et al.[6]
		Kim et al.[13]
		Lee et al.[16]

상대적인 노출 위험도를 산출했다. 하지만 개인정보의 실질적인 노출 가능성을 파악하기 위해서는 웹사이트의 구조적인 부분의 고려가 반영될 필요가 있다.

추가적으로 법률적 내용에 기반 하여 기업 또는 기관이 스스로 개인정보보호 수준을 측정할 수 있는 지표를 제안하는 Kim[11], Kim et al.[12], Lee and Lee[17], Shin et al.[23] 연구와 개인정보 유출에 의한 피해 규모를 산출하기 위한 방법을 제안한 Han et al.[7], Kim[10]의 연구가 있다. 그러나 위의 연구는 기업 및 기관을 대상으로 개인정보보호법 기반의 개인정보 수준 측정 점검 모델이나 피해 규모 산출을 위한 평가모델을 제시하며, 실제 웹상의 개인정보 노출 시를 고려한 위험도가 아니기 때문에 우리가 정의하고자하는 위험도와는 거리가 있다.

3. 개인정보 노출 위험도 계산 모델

본 장에서는 개인정보 노출 위험도 계산 모델을 제시한다. 제 3.1절에서는 개인정보의 위험수준을 정의하고, 제 3.2절에서는 개인정보의 노출 정도 파악을 위한 웹페이지 예상 방문 확률을 계산한다. 이를 바탕으로 제 3.3절에서는 웹사이트 개인정보 노출 위험도 계산 모델을 제시한다.

3.1 개인정보 위험수준

상이한 민감도를 가지는 개인정보들은 차등적으로 위험수준을 달리해야 한다. 예를 들면, 신용카드번호, 주민등록번호의 경우 민감도가 높으며 전화번호, E-Mail 주소는 상대적으로 민감도가 낮다.

본 연구에서 개인정보 위험수준은 기 연구된 개인정보 영향도를 이용하여 개인정보의 민감도에 따른 위험정도를 고려한다[9, 19]. 또한 개인정보의 항목은 안전행정부의 '공공기관 홈페이지 개인정보 노출방지 가이드라인'에서 제시한 개인정보 종류 이외에도 노출될 시 위험하다고 생각되는 전화번호, E-Mail 주소, IP 주소, 사업자등록번호, 법인등록번호를 추가하여 총 12가지의 개인정보를 대상으로 한다[18].

<Table 2>은 민감도에 따른 개인정보의 위험수준을 보여준다. ℓ_3의 개인정보는 민감도가 높은 개인정보로 노출되어 악용될 경우 금전적인 피해가 발생할 수 있지만, ℓ_1의 개인정보는 어느 정도 공개가 되어 있고 개인의 신분이나 신상정보를 파악하기는 어렵다.

〈Table 2〉 Risk Levels of Privacy Information According to Sensitivity

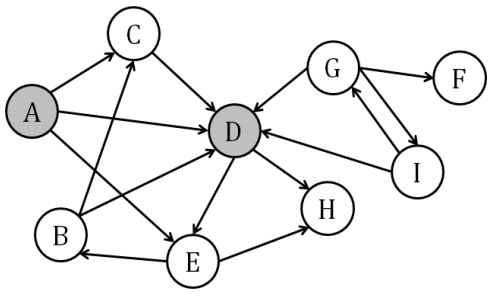
Risk levels	Degrees of sensitivity	Privacy information
ℓ_3	This level contains the most sensitive information. In case of exposure to a third party, the financial damage can occur.	Credit Card Number, Resident Registration Number
ℓ_2	This level contains the sensitive information. In case of exposure, it possibly occurs additional privacy information theft.	Drivers License Number, Passport Number, Account Number, Health Insurance Number
ℓ_1	This level contains some publicly available information. But, in case of exposure, it is concerned about invading their privacy.	Phone Number, Telephone Number, E-Mail Address, IP Address, Business License Number, Corporation Registration Number

3.2 웹페이지의 예상 방문 확률

웹페이지에 따른 개인정보의 노출 정도 파악을 위한 웹페이지 예상 방문 확률의 측정은 링크 기반 구조를 이용한다. 링크는 웹사이트 내 웹페이지 간 이동 및 내비게이션을 이루게 하는 웹사이트의 이용성을 측정하는 요소로 활용될 수 있으며[1], 방향성이 존재하기 때문에 모든 웹페이지들은 특정 웹페이지를 가리키는 링크인 백 링크(Back Link)와 특정 웹페이지에서 밖으로 나가는 링크인 포워드 링크(Forward Link)를 갖는다[2]. 사용자들의 방문 행태는 웹페이지를 임의로 방문하며 탐색함을 가정하며 웹페이지 간 이동은 링크의 클릭으로 이루어진다. 백 링크를 많이 받는 웹페이지는 그 웹페이지로의 이동 경로가 증가하는 것을 의미하며 상대적으로 더 높은 방문 확률을 갖게 된다.

또한 한 웹페이지의 방문 확률이 높을수록 포워드 링크가 가리키는 웹페이지 방문 확률 역시 더 커진다. 예를 들어 <Figure 1>과 같이 웹페이지 A가 웹페이지 C, D, E로 총 3개의 링크가 존재한다면, A의 방문자 중 1/3이 D로 전이된다. 이와 같은 방법으로 웹페이지 간 방문 확률 값 계산을 반복하여 각 웹페이지별 최종 방문 확률을 구한다.

본 연구에서는 이를 측정하기 위해 HITS 알고리즘을 사용한다. HITS 알고리즘은 웹과 같은 링크 구조를 가지는 문서에 상대적 중요도에 따라 가중치를 부여하는 방법이다. 특정 웹페이지의 중요도는 백 링크에 연결된 웹페이지의 중요도 합으로 계산되며, 포워드 링크에 연결된 웹페이지 중요도에 영향을 끼친다. 따라서 연속적으로 파급되는 형태로 확률 값이



〈Figure 1〉 Example of a Website's Structure Consisting of Webpages Connected via Links

전이되기 때문에 반복적인 방법(Iterative Method)으로 계산해야 한다[8, 14].

HITS 알고리즘은 링크의 빈도수와 백 링크에 연결된 노드의 가중치를 고려하여 중요도를 부여하는 원리이기 때문에 웹페이지 방문 확률 계산과 똑같은 원리이다. 또한, 이 알고리즘은 정해진 노드들의 집합 내에서 각 노드의 중요도를 출력하며, 본 논문에서는 정해진 웹사이트 내의 웹페이지들의 집합에서 각 웹페이지의 방문 확률을 출력한다.

구체적으로, 웹페이지 p_{ij}가 가지는 방문 확률은 식 (1)과 같이 계산한다(<Table 3>은 제안된 개인정보 노출 위험도 계산 모델에서 사용되는 기호를 정의하고 있다).

$$e(p_{ij}) = \sum_{p_{ij'} \in B(p_{ij})} \frac{e(p_{ij'})}{N_F(p_{ij'})}, \quad i = 1, 2, \cdots, N, \quad j = 1, 2, \cdots, N(p_i). \tag{1}$$

특정 웹페이지가 가지는 예상 방문 확률은 백 링크들의 웹페이지 예상 방문 확률을 백 링크의 포워드 링크의 개수로 나눈 값의 합으로

정의된다. 즉, p_{ij}의 예상 방문 확률 $e(p_{ij})$는 식 (1)과 같이 p_{ij}의 백 링크인 $p_{ij'}$의 예상 방문 확률 $e(p_{ij'})$를 $p_{ij'}$의 포워드 링크의 개수를 의미하는 $N_F(p_{ij'})$로 나눈 후, 모두 합한 값이다. 예를 들어 $e(p_{ij'})$가 0.3이고, $p_{ij'}$의 포워드 링크의 개수가 3개이면 p_{ij}를 방문할 확률이 1/3로 감소하기 때문에 $e(p_{ij})$는 0.1로 계산된다.

사용자들의 웹페이지 방문 행태는 링크 참조를 통해 이루어져 연속적으로 영향을 주기 때문에 반복적인 계산을 수행한다. 여기서 $e^{(z)}(p_{ij})$는 z번 반복한 후의 p_{ij}의 웹페이지 방문 확률을 의미한다. 초기의 모든 웹페이지의 예상 방문 확률 $e^{(0)}(p_{ij})$, $\forall i, j$는 특정 상수로 동일하다고 가정한다. 이를 초깃값으로 하여 $e^{(1)}(p_{ij})$부터는 식 (2)와 같이 계산한다.

〈Table 3〉 Notations Used in the Proposed Model

Notations	Descriptions
p_i	The i th website
p_{ij}	The j th webpage of the i th website
$R(p_i)$	The risk of p_i
$r(p_{ij})$	The risk of p_{ij}
L	The number of risk levels of privacy information
N	The number of collected websites
$N(p_i)$	The number of webpages in p_i
$N_F(p_{ij})$	The number of forward links in p_{ij}
$e(p_{ij})$	The visiting probability of p_{ij}
$V(p_i)$	The average number of daily visits of p_i
$v(p_{ij})$	The average number of visits of p_{ij}
$C_k(p_{ij})$	The number of privacy information in ℓ_k exposed to p_{ij}
$B(p_{ij})$	The set of back links in p_{ij}

$$e^{(z)}(p_{ij}) = \sum_{p_{ij'} \in B(p_{ij})} \frac{e^{(z-1)}(p_{ij'})}{N_F(p_{ij'})}, \quad i = 1, 2, \cdots, N, \quad j = 1, 2, \cdots, N(p_i). \tag{2}$$

식 (2)에서 $e^{(z)}(p_{ij})$을 계산할 때 바로 전단계의 계산 결과인 $e^{(z-1)}(p_{ij})$을 활용하여 웹페이지 p_{ij}의 예상 방문 확률이 일정한 값으로 수렴할 때까지 반복 수행한다.

3.3 웹사이트 위험도 계산 모델

각각의 웹페이지의 방문자 수의 실제 측정된 값은 얻기 어려우므로 해당 웹사이트의 전체 방문자 수를 이용하여 기댓값을 계산한다. 웹페이지 p_{ij}의 방문자 수의 기대값은 식 (3)과 같이 제 3.2절에서 구한 웹페이지 p_{ij} 예상 방문 확률과 실제 관측치인 웹사이트 p_i의 평균 방문자 수의 곱으로 표현된다.

$$v(p_{ij}) = V(p_i) \cdot e(p_{ij}), \quad i = 1, 2, \cdots, N, \quad j = 1, 2, \cdots, N(p_i). \tag{3}$$

또한, 웹페이지 p_{ij}의 위험도는 식 (4)와 같이 웹페이지 p_{ij}의 방문자 수와 웹페이지 p_{ij}에 노출된 개인정보의 개수에 개인정보 위험수준의 가중치를 부여하여 모두 합한 것과의 곱으로 계산된다. 식 (4)에서 $\sigma(\ell_k)$는 개인정보의 위험수준 ℓ_k의 가중치를 의미한다.

$$r(p_{ij}) = v(p_{ij}) \cdot \sum_{k=1}^{L} \{\sigma(\ell_k) \cdot C_k(p_{ij})\}, \quad i = 1, 2, \cdots, N, \quad j = 1, 2, \cdots, N(p_i). \tag{4}$$

식 (4)의 $\sigma(\ell_k)\cdot C_k(p_{ij})$를 통해 개인정보 위험수준을 고려한 개인정보 위험도가 도출되고, 여기에 웹페이지 p_{ij}의 방문자 수인 $v(p_{ij})$를 곱하게 되면 해당 웹페이지에 노출된 개인정보의 위험수준과 사용자에게 노출될 확률까지 고려된 웹페이지 p_{ij}의 위험도 $r(p_{ij})$가 계산된다.

따라서 개인정보 위험도가 아무리 높아도 웹페이지의 예상 방문 확률이 낮다면 노출될 가능성이 작아지므로 웹페이지 위험도가 낮아지고, 반대로 개인정보 위험도가 낮음에도 불구하고 웹페이지의 예상 방문 확률이 높으면 하나의 개인정보라도 노출될 가능성이 커지기 때문에 웹페이지 위험도가 높아진다.

$$R(p_i)=\sum_{j=1}^{N(p_i)} r(p_{ij}),\quad i=1,\ 2,\cdots,\ N,\quad j=1,\ 2,\cdots,\ N(p_i). \tag{5}$$

최종적으로 웹사이트 p_i의 위험도는 식 (5)와 같이 웹사이트 p_i 내의 모든 웹페이지들의 위험도인 $r(p_{ij}),\ j=1,2,\cdots,N(p_{ij})$의 합으로 계산된다.

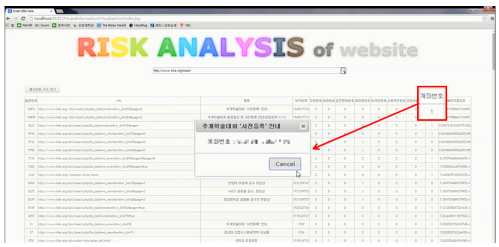
4. 실험 및 결과

4.1 시스템 구현

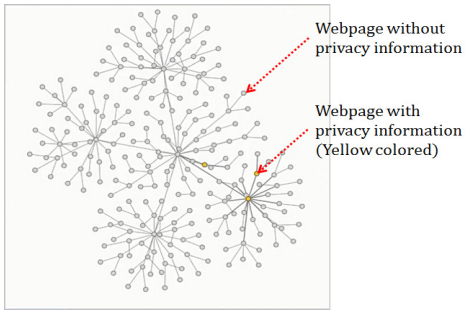
본 절에서는 제안한 개인정보 노출 위험도 계산 모델을 적용하여 구현한 시스템을 설명한다.

<Figure 2>는 특정 웹사이트에서 개인정보가 노출된 웹페이지의 URL, 제목, 노출된 개인정보의 개수 그리고 웹페이지의 예상 방문 확률 순으로 출력된다. 각 노출된 개인정보의 개수를 클릭하면 노출된 실제 개인정보가 나타난다.

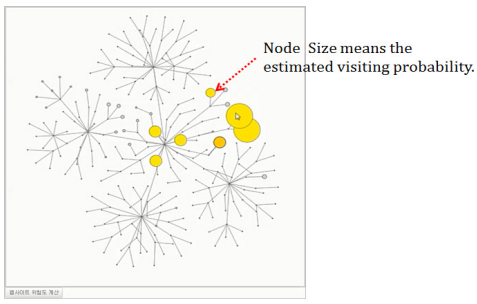
<Figure 3>은 수집된 웹사이트의 웹페이지 구조를 시각화하고 있다. 각각의 노드는 웹페



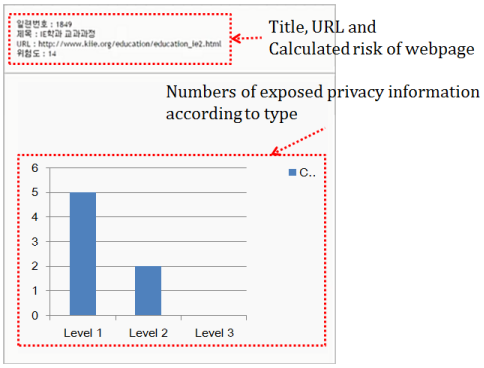
〈Figure 2〉 Exposure Status of Privacy Information in a Website



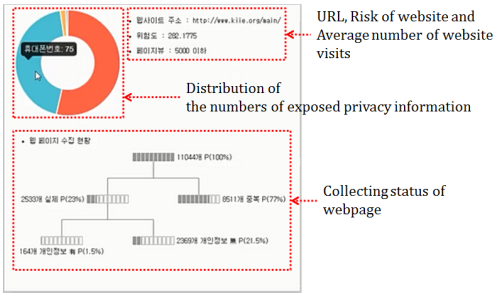
〈Figure 3〉 Visualization of Website's Structure



〈Figure 4〉 Visualization of the Website's Structure Based on the Visiting Probability of Webpages



〈Figure 5〉 Example of Privacy Information in a Webpage



〈Figure 6〉 Distribution of Privacy Information Exposed to the Website and Collecting Status of Webpages

이지를 의미하며, 노란색의 경우 개인정보가 노출된 웹페이지를 의미한다.

다음으로, <Figure 4>는 <Figure 3>에서 출력된 그래프에 각각의 웹페이지에 예상 방문 확률을 부여한 결과를 보여준다. 노드의 크기는 웹페이지의 예상 방문 확률의 크기를 의미한다.

각 웹페이지에 대해서, <Figure 5>와 같이 Title, URL, 웹페이지 위험도를 확인할 수 있다. 또한, 개인정보의 노출 개수가 위험수준별로 그래프에 표현된다.

<Figure 6>에서는 웹사이트의 도메인 주소와 일평균 방문자 수, 웹사이트 위험도, 도넛형 차트로 표현한 개인정보 노출 분포, 웹페이지 수집 상태를 확인할 수 있다.

4.2 실험 환경

본 연구에서 제안한 개인정보 노출 위험도 계산 모델을 평가하기 위해 실제 웹사이트를 수집하여 구성한 데이터 세트를 사용하였다. 데이터 세트는 크게 4개의 분류로 구성된다. 구체적으로는 61개의 대학교 웹사이트와 16개의 은행 웹사이트, 38개의 중앙 행정 기관 웹사이트, 13개의 시・도 교육청 웹사이트를 포함하였다. 웹사이트 선정과 웹사이트의 일평균 방문자 수 데이터는 웹 트래픽 전문 조사기관에서 제공하는 해당 분류별 정보를 이용하였다[22]. 4개의 분류에 대한 설명을 용이하게 하기 위해 대학교는 *U*, 은행은 *B*, 중앙 행정 기관은 *O*, 시・도 교육청은 *M*으로 명명한다. 또한, 개인정보의 인식은 안전행정부의 지침 범위에 따라 개인정보 정규표현식을 활용했다[18]. 이에 따라 정규표현식에 일치하지 않는 경우 또는 텍스트가 아닌 형태로 존재하는 개인정보의 검출오류가 존재함을 알려둔다.

실험의 용이성을 위해서 데이터 수집 및 실험범위를 외부 웹페이지를 가리키는 링크는 고려하지 않고, 각 웹사이트 내의 같은 도메인을 가진 웹페이지들의 관계로 한정하였다. 또한, 최초 웹페이지를 기준으로 특정 웹페이지에 도달하는 데 필요한 “클릭”의 수를 의미하는 링크의 깊이를 5로 설정하고, 한 웹 서버에서 수집할 최대 웹페이지 수는 2,000개로 설정하였다. 웹페이지 예상 방문 확률 계산 시 계산 반복수는 25번으로 설정하였으며, 수집한 웹사이트는 총 128개였다.

그리고 개인정보 위험수준의 가중치를 의미하는 $\sigma(\ell_1)$, $\sigma(\ell_2)$, $\sigma(\ell_3)$는 편의상 각각 1, 2, 3으로 설정하여 가중치 값에 따른 큰 차이는 없도록 하였다. 실 환경에서의 분석 시에는 항목 간의 상대적 중요도 정책에 따른 가중치가 조사되어 설정되어야 함을 알려둔다.

4.3 실험 결과 및 분석

<Table 4>는 분류별 평균 웹사이트 위험도를 보여준다. 은행은 평균 위험도가 가장 높았고 시·도 교육청은 가장 낮음을 알 수 있다. 다른 분류에 비해 은행의 높은 평균 위험도를 분석해보니 위험도의 표준편차가 매우 높았고 특정 2개의 웹사이트가 매우 높은 방문자 수를 가지고 있음이 밝혀졌다. 그래서 은행의 웹사이트 중 특정 2개의 웹사이트를 제외한 B^*의 기초통계량 결과를 보니 다른 분류에 비해 낮은 평균 위험도를 가졌고 상대적으로 잘 관리되고 있음을 확인할 수 있었다.

<Table 5>는 분류별 일평균 방문자 수 기준 상위 5개의 웹사이트에 대한 일평균 방문자 수와 분석된 웹사이트 위험도를 보여준다. 웹사이트의 일평균 방문자 수는 은행이 다른 분류에 비해 급격히 높음을 확인할 수 있었다.

또한, 웹사이트 *B1*의 경우 일평균 방문자 수가 매우 높지만 노출된 개인정보가 적기 때문에 위험도가 매우 낮게 계산되었다. 그리고 웹사이트 *M3*의 경우 일평균 방문자 수도 낮을 뿐만 아니라 노출된 개인정보가 상대적으로 웹페이지 방문자 수가 적은 곳에 분포하여 위험도가 매우 낮게 계산되었다. 웹사이트 구조상, 웹사이트의 전체 사용자가 방문하는 웹페이지는 소수에 불과하기 때문에 웹페이지 간의 방문자 수의 격차는 매우 커지게 된다. 이는 해당 웹페이지 위험도 계산에 영향을 끼치기 때문에 매우 낮은 위험도 값이 존재하게 된다.

〈Table 4〉 Basic Statistics of Website Risk According to Category

Category	Min	Max	Mean	Standard deviation
U	0	992,015	68,063	151,813
B	**0**	**886,516**	**115,976**	**241,260**
B*	**0**	**333,329**	**40,921**	**96,863**
O	0	707,207	73,306	168,402
M	0	67,440	16,308	27,585

B^* : The bank websites considered without the most two websites in terms of website visits.

〈Table 5〉 Top 5 Websites According to Category Based on Their Average Number of Daily Visits

Categories	List of websites	Average numbers of daily visits	Website risks
U	*U1*	647,986	992,015
	U2	472,870	470,590
	U3	177,747	17,292
	U4	158,923	138,536
	U5	150,346	338,069
B	*B1*	38,057,238	5.3E-15
	B2	16,149,621	175,934
	B3	15,611,989	396,201
	B4	5,697,238	886,519
	B5	3,629,575	333,329
O	*O1*	1,863,701	694,561
	O2	646,270	272,786
	O3	557,978	707,207
	O4	404,789	15,495
	O5	329,521	56,271
M	*M1*	70,612	59,343
	M2	28,476	1,758
	M3	15,812	1.2E-06
	M4	9,532	66,067
	M5	9,157	67,440

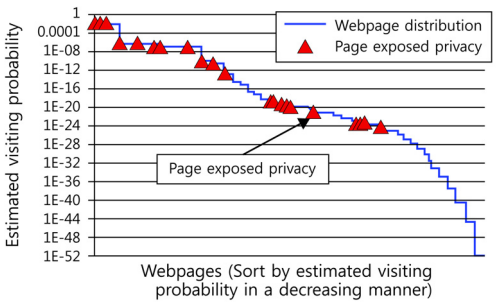
〈Table 6〉 Comparison of Websites *W1* and *W2*

Website	Number of exposed privacy information	Average number of daily visits	Website risk
W1	3,299	6,340	67,529.9
W2	13,498	12,382	37,159.7

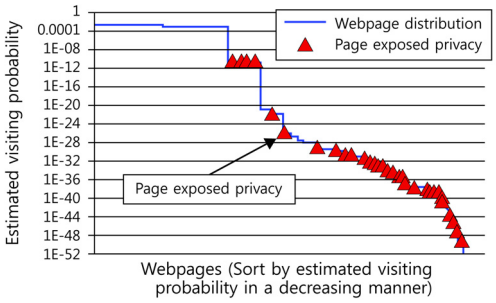
다음은 웹페이지의 예상 방문 확률의 고려가 위험도 계산에 미치는 영향을 살펴보기 위해 전체 웹사이트에서 임의의 웹사이트 *W1*와 웹사이트 *W2*를 선정하여 분석하였다. <Figure 7>과 <Figure 8>에서 웹사이트 *W1*는 웹사이트 *W2*에 비해서 웹페이지의 예상 방문 확률이 높은 쪽에 많은 개인정보가 노출된 것을 확인할 수 있다.

<Table 6>을 보면, 웹사이트 *W2*에 노출된 개인정보의 개수와 웹사이트 평균 방문자 수가 웹사이트 *W1*에 비해 2~4배 정도 높음에도 불구하고, 웹사이트 위험도가 낮다. 그 이유는 웹사이트 *W1*의 개인정보가 노출된 웹페이지가 웹사이트 *W2*에 비해 웹페이지 예상 방문 확률이 높은 곳에 분포되어 있기 때문이다 (<Figure 7> 참고). 즉, 웹사이트 *W1*에 노출된 개인정보가 사용자에게 노출될 가능성이 더 큰 것으로 볼 수 있다.

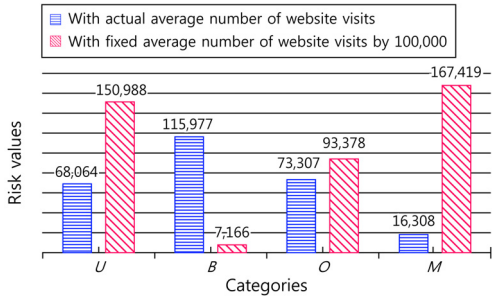
다음으로는, 웹사이트 평균 방문자 수는 변동 가능성이 큰 요소이기 때문에 평균 방문자 수를 고정하고 웹사이트 위험도를 분석하여, 웹사이트 방문자 수와 독립적으로 각 분류별 개인정보 노출관리 상황을 살펴보았다. <Figure 9>는 웹사이트의 평균 방문자 수를 100,000으로 동일하다고 가정한 후, 웹사이트 위험도를 나타낸다. 은행을 제외한 3개 분류의 위험도가 은행보다 13배 이상 높아진다. 이는 은행 카테고리의 웹사이들은 개인정보 노출관리 측면에서 우수하나, 다른 카테고리의 웹사이트들에 비해 평균 방문자 수가 매우 높기 때문에 소수의 개인정보라도 큰 위험도를 야기하고 있음을 의미한다.



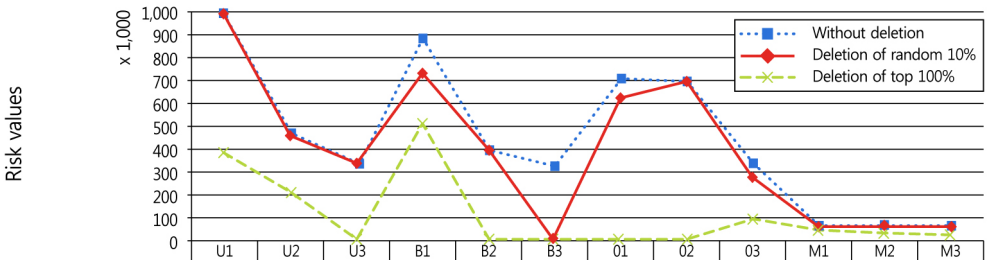
〈Figure 7〉 Distribution of Webpages in Website *W1*



〈Figure 8〉 Distribution of Webpages in Website *W2*



〈Figure 9〉 Comparison of Website Risks According to Website Visits



	U1	U2	U3	B1	B2	B3	01	02	03	M1	M2	M3
Without deletion	992,015	470,590	338,069	886,519	396,202	333,329	707,207	694,562	337,057	67,440	66,067	59,343
Deletion of random 10%	992,013	452,525	331,261	731,587	396,196	2.07E-05	622,860	694,560	280,803	49,159	57,047	56,220
Deletion of top 100%	379,835	209,521	4	507,799	9.97E-12	7.38E-09	19	14	90,476	42,540	33,238	24,695

〈Figure 10〉 Comparison of the Efficiency of Priority-based Removal for Privacy Information

마지막으로 <Figure 10>은 노출된 개인정보의 우선순위 기반 삭제의 효율성을 판단하기 위해 삭제 없이 계산된 웹사이트 위험도와 개인정보 10%를 무작위로 삭제한 후의 웹사이트 위험도, 개인정보 위험도 계산 결과의 상위 10%의 개인정보를 삭제한 후의 웹사이트 위험도를 비교한다.

무작위 삭제보다 우선순위 삭제가 웹사이트의 위험도를 약 60% 이상 감소시킨다. 따라서 본 연구에서 제안하는 개인정보 위험도에 따른 우선순위 삭제를 실행한다면 해당 웹사이트의 개인정보 노출 가능성이 효율적으로 감소할 수 있게 된다.

5. 결 론

본 논문에서 개인정보의 민감도에 따른 위험수준과 웹페이지의 구조적 특성을 고려한 노출확률을 반영하여 웹사이트 개인정보 노출 위험도를 도출하였다. 또한, 자동화 시스템을 구현하여 각 분류별 웹사이트 위험도를 비교 분석함으로써 개인정보의 우선순위별 대처가 가능한 합리적인 기준 제안을 목적으로 한다.

한국의 대학교, 은행, 중앙 행정 기관, 시・도 교육청 4가지 분류의 웹사이트 위험도를 측정해보니 시・도 교육청, 중앙 행정 기관, 대학교의 경우, 은행에 비해 상대적으로 노출 가능성은 작으나 웹사이트 위험도가 높게 측정되었다. 또한, 은행의 경우 개인정보 노출에 대한 관리는 잘되고 있었으나 사용자가 방문할 확률이 매우 높아 소수의 개인정보 노출이라도 그 위험성이 다른 분류에 비해 매우 높음을 확인할 수 있었다.

이를 바탕으로 본 연구는 다음과 같은 실용적 가치를 가진다. 첫째, 웹사이트 사용자들에게 개인정보 노출에 대한 태도의 변화를 촉구하여 사용자 스스로도 민감한 정보를 단속하고 웹사이트 내 개인정보 입력 시 신중하게 판단하는 등 자기정보보안 의식을 향상에 기여할 수 있다. 둘째로, 웹사이트 운영자들이 개인정보 노출에 대한 모니터링 가능 및 우선순위 대처가 가능하도록 판단 지표를 제공한다. 마지막으로 포괄적인 관점에서는 정부에서 개인정보 노출 관련 정책 수립 시, 웹사이트 내 노출된 개인정보의 산업 군별 분류를 통해 군별 대처방안의 기초자

료로 활용할 수 있을 것으로 기대한다.

추가로 산업 군별 노출실태 분석 및 시간에 따른 개인정보의 노출추세 파악을 위해 더욱 포괄적인 범위의 데이터 세트를 구성하여 지속적으로 연구를 진행할 필요가 있다. 또한, 웹사이트 내에 비정형 개인정보가 다수 존재함에도 불구하고 정규표현식에 일치하지 않는 개인정보나 텍스트가 아닌 형태로 존재하는 개인정보의 검출오류가 존재하므로 다양한 형식에 포함된 개인정보 식별을 위해 정교화된 개인정보 검출 및 식별에 대한 추가 연구가 필요하다.

References

[1] BizSpring Education Consulting Team, “Website Measurement and Analysis,” BizSpring, p. 87, 2011.

[2] Brin, S. and Page, L., “The Anatomy of a Large-scale Hypertextual Web Search Engine,” Journal of Computer Networks and ISDN Systems, Vol. 30, No. 1-7, pp. 107-117, 1998.

[3] Cheon, M. H., Choi, J. S., and Shin, Y. T., “Measuring Method of Personal Information Leaking Risk Factor to Prevent Leak of Personal Information in SNS,” Journal of the Korean Institute of Information Security and Cryptology, Vol. 23, No. 6, pp. 1199-1206, 2013.

[4] Cho, S. and Jun, M., “Privacy Leakage Monitoring System Design for Privacy Protection,” Journal of the Korean Institute of Information Security and Cryptology, Vol. 22, No. 1, pp. 99-106, 2012.

[5] Choi, D. S., Kim, S. H., Jo, J. M., and Jin, S. H., “Big Data Privacy Risk Analysis Technique,” Korea Institute of Information Security and Cryptology Review, Vol. 23, No. 3, pp. 56-60, 2013.

[6] Choi, J. Y., Ha, T. G., Lee, G. S., and Won, Y. J., “Privacy Incident Response System,” Journal of the Korea Institute of Information Security and Cryptology, Vol. 19, No. 6, pp. 9-14, 2009.

[7] Han, C. H., Chai, S. W., Yoo, B. J., Ahn, D. H., and Park, C. H., “A Quantitative Assessment Model of Private Information Breach,” The Journal of Society for e-Business Studies, Vol. 16, No. 4, pp. 17-31, 2011.

[8] Kim, B. M., Han, S. Y., and Kim, Y. C., “Design of Advanced HITS Algorithm by Suitability for Importance Evaluation of Web-Documents,” The Journal of Society for e-Business Studies, Vol. 8, No. 2, pp. 23-31, 2003.

[9] Kim, E., “Privacy Detection and Risk Analysis Model,” Master’s Theses for Graduate School of Sungshin Woman’s University, 2010.

[10] Kim, J. Y., “Analyzing Effects on Firms’ Market Value of Personal Information Security Breaches,” The Journal of Society for e-Business Studies, Vol. 18, No. 1, pp. 1-12, 2013.

[11] Kim, M. S., “The Study of Check-list

Based on Privacy Law in Korea for Private Company," Proceedings of the Korean Information Science Society 2010 Conference, Vol. 37, No. 2(B), pp. 37-42, 2010.

[12] Kim, M. S., Noh, B. N., and Kim Y. M., "A Privacy Level Check Model Based on New Privacy Law in Korea," Proceedings of the Korean Information Science Society 2011 Conference, Vol. 35, No. 1(D), pp. 118-121, 2011.

[13] Kim, P., Lee, Y. H., and Khudaybergenov, T., "A Method for Quantitative Measuring the Degree of Damage by Personal Information Leakage," Journal of the Korean Institute of Information Security and Cryptology, Vol. 25, No. 2, pp. 395-410, 2015.

[14] Kleinderg, J., "Authoritative Sources in a Hyperlinked Environment," Journal of the ACM, Vol. 46, No. 5, pp. 604-632, 1999.

[15] Lee, G. H. and Young, J. D., "A Study of Measurement Methods and Practical Cases on Leakage Risk of Privacy Information in Private Sector," Journal of the Korean Institute of Information Security and Cryptology, Vol. 18, No. 3, pp. 92-100, 2008.

[16] Lee, K. S., Ahn, H. B., and Lee, S. Y., "A Study on a Prevention Method for Personal Information Exposure," Journal of Information and Security, Vol. 12, No. 1, pp. 71-77, 2012.

[17] Lee, S. J. and Lee, Y. J., "Development of a New Instrument to Measuring Concerns for Corporate Information Privacy Management," Journal of Information Technology Applications and Management, Vol. 16, No. 4, pp. 79-92, 2009.

[18] Ministry of Government Administration and Home Affairs, "Homepage Personal Information Exposure Guidelines," p. 35, 2014.

[19] Ministry of Public Administration and Security (MOSPA)-Korea Internet and Security Agency (KISA), "Perform Manual of Privacy Impact Assessment in Public Authorities," pp. 78-81, 2015.

[20] Nevermind, "Principal of Long tail, Pareto and Short tail," [URL] http://nevermind.tistory.com/2.

[21] Park, S. J. and Lim, J. I., "A Study on the Development of SRI(Security Risk Indicator)-Based Monitoring System to Prevent the Leakage of Personally Identifiable Information," Journal of The Korea Institute of Information Security and Cryptology, Vol. 22, No. 3, pp. 637-644, 2012.

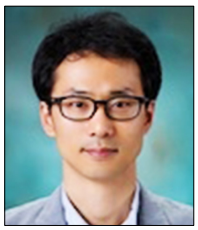
[22] Ranky.com, "Professional Website Analysis/Evaluation Organization-Webpage View During Oct. 01-07, 2014," [URL] http://www.rankey.com/.

[23] Shin, Y. J., Jeong, H. C., and Kang, W. Y., "A Study of Priority for Policy Implement of Personal Information Security in Public Sector: Focused on Personal Information Security Index," Journal of the Korean Institute of Information Security and Cryptology, Vol. 22, No. 2, pp. 379-390, 2012.

저 자 소 개

이수경 (E-mail: dltnrud1212@naver.com)
2012년 인천대학교 산업경영공학과 (학사)
2016년~현재 인천대학교 산업경영공학과 (석사과정)
관심분야 개인정보 위험도

손진식 (E-mail: realrice@kepco-enc.com)
2015년 인천대학교 산업경영공학과 (학사)
2016년~현재 한국전력기술(주) 기술원



김관호 (E-mail: khokim@inu.ac.kr)
2006년 동국대학교 정보시스템전공 (학사)
2012년 서울대학교 산업공학과 (박사)
2013년 경희대학교 연구박사
2014년~현재 인천대학교 산업경영공학과 조교수
관심분야 통계적 기계학습, 빅데이터