JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 22, No. 1 : 69~74 Jan. 2018

웹 문서 분석 기반 개인정보 위험도 분류 시스템

이형선 · 임재돈 · 정회경*

Web Document Analysis based Personal Information Hazard Classification System

Hyoungseon Lee · Jaedon Lim · Hoekyung Jung*

Department of Computer Engineering, Paichai University, Daejeon 35345, Korea

요 약

최근 개인정보 유출로 인해 피싱이나 스팸 등이 발생하고 있다. 기존에 시스템들은 개인정보 유출의 사전 예방에 중점을 두고 있다. 이로 인해 이미 유출된 개인정보가 있을 경우 개인정보 유출판별을 할 수 없는 문제점이 있었다. 이에 본 논문에서는 웹 문서 분석 기반 개인정보 위험도 분류 시스템을 제안한다. 이를 위해 트위터 서버로부터 웹 문서를 수집하고 해당 웹 문서 내에 사용자가 입력한 검색어가 있는지 확인한다. 또한 웹 문서 내에 유출된 개인정보들의 위험도 분류 가중치를 계산하고 개인정보를 유포한 트위터 계정의 권위를 확인한다. 이를 기반으로 위험도를 도출하여 해당 웹 문서의 개인정보 유출여부 판별을 확인할 수 있다.

ABSTRACT

Recently, personal information leakage has caused phishing and spam. Previously developed systems focus on preventing personal information leakage. Therefore, there is a problem that the leakage of personal information can not be discriminated if there is already leaked personal information. In this paper, we propose a personal information hazard classification system based on web document analysis that calculates the hazard. The system collects web documents from the Twitter server and checks whether there are any user-entered search terms in the web documents. And we calculate the hazard classification weighting of the personal information leaked in the web documents and confirm the authority of the Twitter account that distributed the personal information. Based on this, the hazard can be derived and the user can be informed of the leakage of personal information of the web document.

키워드 : 개인정보, 수집, 웹 문서, 유출 판별

Key word : Collection, Leak Detection, Personal Information, Web Document

Received 18 December 2017, **Revised** 24 December 2017, **Accepted** 31 December 2017

* **Corresponding Author** Hoekyung Jung(E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Department of Computer Engineering, Paichai University, Daejeon 35345, Korea

Open Access http://doi.org/10.6109/jkiice.2018.22.1.69 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

소셜 네트워크 서비스는 사진, 사생활, 전화번호 등의 수많은 자료들이 검색, 변조, 악용, 스팸 등으로 악용되고 있다. 국내에서 발생한 개인정보 유출 사례는 대표적으로 2011년 7월 26일 유명 포털사이트인 네이트와 싸이월드가 해킹되어 3,500만 명의 개인정보가 유출되었다[1,2]. 이를 통해 개인정보 유출에 대한 문제가 발생하고 있음을 알 수 있다.

기존 시스템들은 특정 검색어가 이메일을 포함되었는지 검색하여 분류한다[3,4]. 또는 특정 웹 사이트에서는 사용자가 해당 웹 사이트의 검색 기능을 이용하여 사용자가 직접 개인정보 유출을 판별해야 했다[5-7]. 이와 같이 개인정보 유출을 판별하는 시스템이나 알고리즘은 개발되지 않은 실정이다.

이에 본 논문에서 제안하는 시스템은 사용자가 어플리케이션을 통해 입력한 검색어가 웹 문서 내에 존재하는지 확인한다. 그리고 사용자가 입력한 검색어가 해당 웹 문서 내에 존재할 경우 위험도 가중치를 계산한다. 위험도 가중치가 임계값 이상일 경우 유포자의 SNS (Social Networking Service) 내 권위를 평가하여 위험도를 도출한다. 도출된 위험도를 기반으로 개인정보 유출여부를 판별한 뒤 사용자에게 FCM(Firebase Cloud Messaging)을 통해 웹 문서의 위험도와 링크를 보여준다.

Ⅱ. 시스템 설계

제안하는 시스템은 개인정보 유출여부를 판별하여 웹 프로그램과 모바일 어플리케이션을 통해 사용자에게 개인정보가 유출된 웹 문서의 위험도와 링크를 보여주는 시스템이다. 시스템은 개인정보 위험도 분류 알고리즘을 기반으로 트위터에서 수집한 웹 문서에 사용자가 입력한 검색어가 있을 경우 가중치를 계산하여 개인정보 유출을 판별한다. 개인정보 유출이 판별 되었을 경우 개인정보 유포자의 트위터 내의 권위를 이용해 위험도를 도출한다. 도출된 위험도를 기반으로 개인정보 유출여부가 판별된 웹 문서를 정렬한 뒤 어플리케이션을 통해 사용자에게 보여준다. 제안하는 시스템의 구성도는 그림 1과 같다.

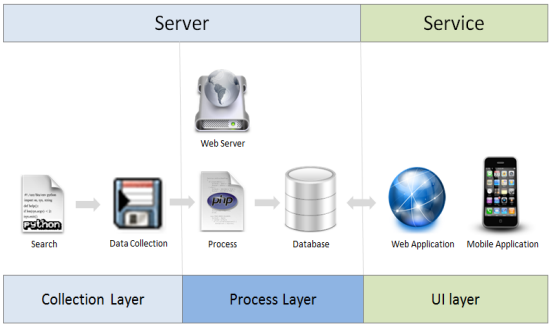

Fig. 1 System Configuration Diagram

서버는 웹 프로그램을 통해 입력된 검색어로 데이터를 수집하고 개인정보 유출을 판별하여 데이터베이스에 적재한다. 사용자는 웹 어플리케이션을 사용하여 검색하고자 하는 검색어를 입력하고 결과를 제공 받을 수 있다. 그림 2는 제안하는 시스템 웹 문서 분석 기반 개인정보 위험도 분류 시스템의 구조도이다.

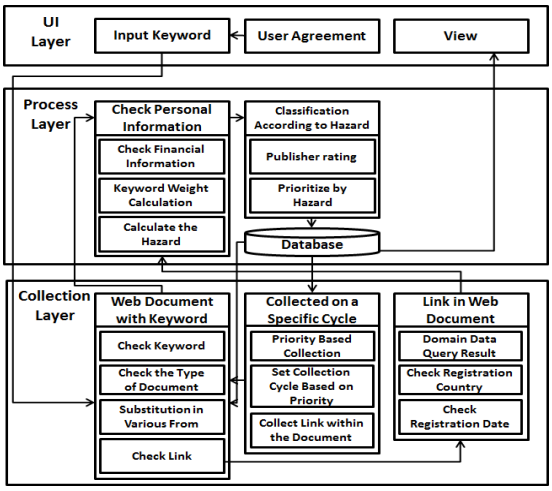

Fig. 2 System Architecture Diagram

수집 계층에서는 어플리케이션을 통해 사용자가 입력한 검색어가 포함되어 있는 웹 문서를 수집하고 해당 링크를 데이터베이스에 적재한다. 처리 계층에서는 수집한 데이터에 위험도 분류 가중치를 적용하여 개인정보 유출여부를 판별하고 계산된 위험도 분류 가중치를 데이터베이스에 적재한다. 사용자 인터페이스 계층에서는 사용자가 검색어를 입력하고 개인정보 수집 이용 약관에 대한 동의를 받는다. 또한 사용자에게 개인정보

가 유출된 웹 문서의 링크와 위험도를 제공해준다. 그림 3은 웹 문서 분석 기반 개인정보 위험도 분류 시스템의 흐름을 나타낸다.

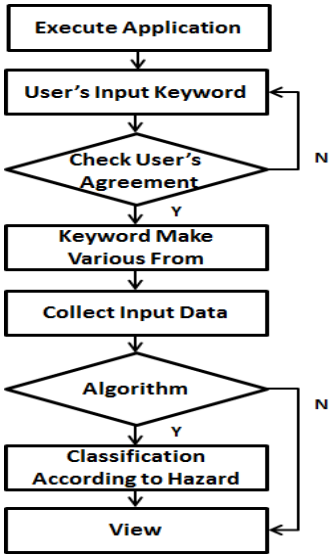

Fig. 3 System Flow Chart

사용자가 어플리케이션을 실행하고 개인정보 이용 및 수집 약관에 동의 할 경우 이름, 전화번호, 주소 등의 검색어를 입력한다. 사용자에게 입력 받은 검색어를 다양한 형태로 검색한다. 예를 들어 02-2134-2134를 입력했을 때 0221342134, 21342134 등의 형태로 검색한다. 또한 검색어가 포함되어 있는 웹 문서를 트위터 서버에 요청을 통해 수집한다. 수집된 웹 문서를 개인정보 위험도 분류 알고리즘을 통해 개인정보 유출 여부 확인과 위험도를 계산한다. 계산된 위험도를 기준으로 웹 문서들을 정렬하여 사용자에게 보여준다.

그림 4는 개인정보 위험도 분류 알고리즘의 흐름을 나타낸다. 수집된 웹 문서에서 검색어 포함 여부를 확인하고 위험도 분류 가중치를 측정한다. 측정된 위험도 가중치를 더하여 가중치가 임계값 이상인지 확인한다. 위험도 분류 가중치가 설정된 임계값 이상일 경우 개인정보 유출로 판별한다. 개인정보 유출로 판별된 웹 문서 내에 있는 트위터 계정의 생성일과 트윗에 유포자의 팔로워 수, 타임라인에 있는 트윗의 수를 기반으로 평가한다. 그리고 평가된 점수와 위험도 분류 가중치를 활용하여 위험도를 계산한 뒤 개인정보가 유출된 웹 문서의 링크와 계산된 위험도를 데이터베이스에 적재한다. 시스템은 데이터베이스에 적재된 링크를 특정 주기마다 다시 수집한다.

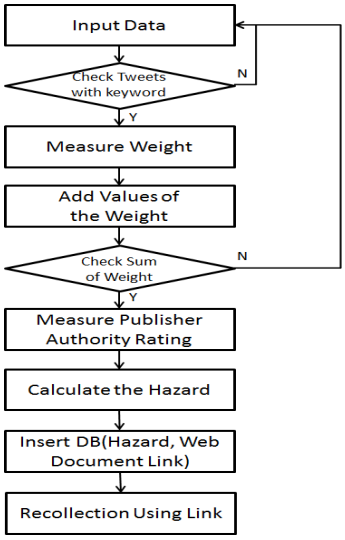

Fig. 4 System Algorithm Flow Chart

제안하는 시스템은 트위터의 URL 규칙성을 활용하여 웹 문서를 수집한다. 데이터 수집 모듈에서는 트위터 서버로 GET 요청을 전송하여 웹 문서를 수집한다. 그림 5는 데이터 수집 모듈의 흐름을 나타낸다.

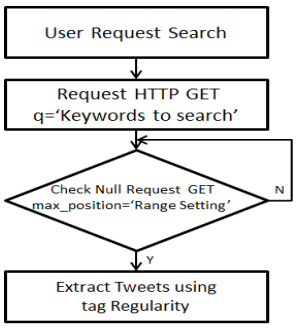

Fig. 5 System Data Module Flow chart

사용자가 웹 프로그램을 통해 검색어를 입력할 시 https://twitter.com.com/search 주소에 사용자가 입력한 검색어를 포함하여 GET 요청을 보낸다. 검색의 범위를 지정하는 인자를 통해 트위터 서버에서 검색어가 포함된 모든 웹 문서를 수집한다.

Ⅲ. 시스템 구현

본 장에서는 제안하는 웹 문서 분석 기반 개인정보 유출 판별 시스템의 구현을 다룬다. 그림 6은 검색하고자 하는 검색어를 입력 받는 페이지이다.

Fig. 6 User Input Searching Keyword Page

성과 이름으로 구별하여 입력하며 주소, 휴대폰 번호, 기타 정보는 ‘,’로 구분하여 입력한다. 기타정보를 기입하지 않아도 다음 페이지로 이동한다. 그러나 성, 이름, 휴대폰 번호, 주소는 입력하지 않을 시 다음 페이지로 이동하지 않는다. 웹 문서에 사용자가 입력한 검색어가 있는지 입력 받은 검색어들을 활용하여 확인한다. 그림 7은 모바일 어플리케이션에서 개인정보가 유출된 웹 문서의 위험도와 링크를 사용자에게 보여주는 페이지이다.

Fig. 7 Person Information Disclosure Hazard Page

유출된 개인정보가 있는 웹 문서의 링크와 위험도를 상, 중, 하 3가지로 분류하여 보여준다. 자연수로 저장된 개인정보 위험도를 표준편차를 사용하여 상, 중, 하 세 가지로 분류한다. 위험도를 기준으로 웹 문서들을 내림차순으로 정렬하여 사용자에게 보여준다.

그림 8은 FCM을 활용하여 푸시 알림을 사용자에게 보여주는 화면을 나타낸다. 개인정보 유출이 있을 경우 FCM을 활용하여 사용자에게 푸시 메시지를 제공한다. 모바일 어플리케이션은 개인정보 유출로 판별된 웹 문서를 확인할 수 있다. 그리고 데이터베이스에 적재된 웹 문서의 위험도와 링크를 기반으로 재수집한 데이터의 알림을 사용자에게 보여준다.

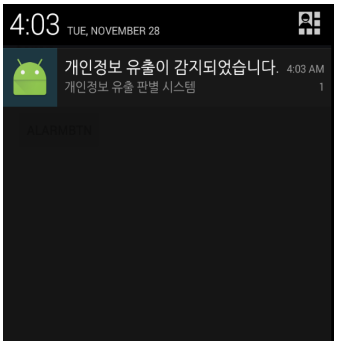

Fig. 8 FCM Message Service

Ⅳ. 고 찰

본 장에서는 기존 시스템과 제안하는 시스템의 실험 및 검증을 다룬다. 제안하는 시스템의 검증을 위해 트위터 서버에서 요청을 통해 수집한 데이터를 통해 수집한 데이터를 활용하여 비교 분석을 진행하였다.

Table. 1 Experiment Data Set

Data Set	Data Type
No Leaked Personal Information	Test A : 1,593 Data Set Test B : 1,523 Data Set
Leaked Personal Information	Test A : 1,890 Data Set Test B : 9,84 Data Set
Amount	Test A : 3,483 Test B : 1.523

표 1은 실험을 위한 데이터를 나타낸다. 실험에 사용한 데이터 셋은 트위터 서버에서 얻은 데이터로 일반 트위터 사용자가 올린 트윗과 실험용 데이터를 사용했다. 그림 9은 실험용 계정을 통해 생성한 2개의 키워드 셋을 기존 시스템과 제안하는 시스템을 통해 비교한 그래프를 나타낸다. 테스트 A에서는 총 웹 문서 3483개

중 개인정보가 유출된 웹 문서가 1890개이다. 기존 시스템은 1237개의 웹 문서를 개인정보 유출로 판별하였고 정확도는 약 65.44%이다. 제안하는 시스템은 1435개의 웹 문서를 개인정보 유출로 판별하였고 정확도는 약 75.92%이다.

테스트 B에서는 총 웹 문서 1523개 중 개인정보가 유출된 웹 문서가 984개이다. 기존 시스템은 527개의 웹 문서를 개인정보 유출로 판별하였고 정확도는 약 53.55%이다. 제안하는 시스템은 611개의 웹 문서를 개인정보 유출로 판별하였고 정확도는 62.09%이다.

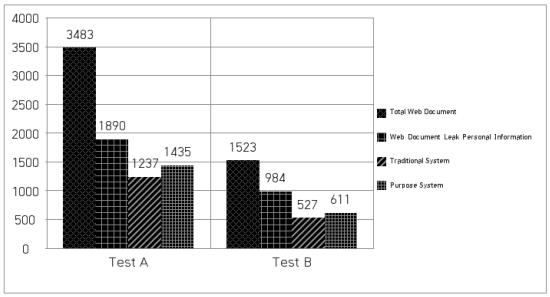

Fig. 9 Data Comparing Graph

두 번의 실험을 분석한 결과 기존 시스템에 비해 제안하는 시스템이 개인정보가 유출된 웹 문서를 판별하는 정확도가 평균 9.51% 향상되었다. 이를 통해 제안하는 시스템이 기존 시스템에 비해 개인정보 유출여부 판별을 보다 정확하게 하는 것을 알 수 있다.

V. 결 론

기존 시스템들은 개인정보 유출을 사전에 예방하기 위해 개발되었다. 그러나 사전 예방에 중점을 두었기 때문에 웹에 유출된 사용자의 개인정보 유출여부를 판별하는 기능을 제공해주지 못하는 문제점이 있다.

이를 해결하기 위해 본 논문에서는 웹 문서 분석 기반 개인정보 위험도 분류 시스템을 제안하였다. 제안하는 시스템은 수집한 웹 문서에서 개인정보 유출여부를 판별한다. 이를 위해 사용자가 입력한 검색어가 있는지 확인한다. 검색어가 있을 경우 시스템은 위험도 분류 가중치들을 계산하여 임계값 이상인지 확인한다. 임계값 이상일 경우 해당 트윗을 게시한 유포자의 권위와 계산한 위험도 분류 가중치를 기반으로 위험도를 측정한다. 그리고 위험도를 기준으로 정렬하여 위험도와 개인정보가 유출된 웹 문서의 링크를 사용자에게 보여준다. 그리고 FCM 푸시 서비스를 이용하여 알림 메시지를 사용자에게 제공한다. 향후 연구로는 다양한 SNS에서도 개인정보 유출 판별 서비스를 제공할 수 있도록 연구를 진행하여야 한다.

ACKNOWLEDGEMENTS

This work was supported by the research grant of Pai Chai University in 2017.

REFERENCES

[1] J. H. Eom, M. J. Kim, “Effect of Information Security Incident on Outcome of Investment by Type of Investors: Case of Personal Information Leakage Incident,” *Journal of The Korea Institute of Information Security & Cryptology*, vol. 26, no. 2, pp. 463-474, Apr. 2016.

[2] Y. H. Kim, I. H. Cho, K. H. Lee, “A Decision-Making Model for Handling Personal Information Using Metadata,” *Journal of The Korea Institute of Information Security & Cryptology*, vol. 26, no. 1, Feb. 2016.

[3] C. S. Lee, Y. H. Kim, “An Analysis of Relationship between Industry Security Education and Capability: Case Centric on Insider Leakage,” *The Journal of Society for e-Business Studies*, vol. 20, no. 2, pp. 27-36, May 2015.

[4] G. H. Song, and K. S. Shim, “Privacy-Preserving Big Data Publication using MapReduce.” *Journal of KIISE*, pp. 175-177, Oct. 2016.

[5] D. Vatsalan, Z. Sehili, P. Christen, E. Rahm, “Privacy-Preserving Record Linkage for Big Data : Current Approaches and Research Challenges,” *Privacy-Preserving Record Linkage for Big Data*, Feb. 2017.

[6] Y. Li, L. Zhou, H. Zhu, “Privacy-Preserving Location Proof for Securing Large-Scale Database-Driven Cognitive Radio Networks,” *IEEE Internet of Things Journal*, vol. 3. no. 4, pp. 563-571, Apr. 2016.

[7] D. W, Park, “Analysis of Mobile Smishing Hacking Trends and Security Measures,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 19, no. 11, pp. 2615-2622, Nov. 2015.

이형선(Hyoungseon Lee)

2016년 배재대학교 컴퓨터공학과(공학사)
2016년 ~ 현재 배재대학교 컴퓨터공학과 석사과정
※관심분야 : 리눅스, 빅데이터 분석

임재돈(Jaedon Lim)

2017년 ~ 현재 배재대학교 컴퓨터공학과 박사과정
2010년 충북대학교 마케팅전공 공학석사
2006년 ~ 2009년 영동대학교 경영학과 조교수
2003년 서원대학교 경영학과 e-비지니스(경영학석사)
2001년 서원대학교 경영정보학과(경영학석사)
※관심분야 : 빅데이터 분석, 재난 예측

정회경(Hoekyung Jung)

1985년 광운대학교 컴퓨터공학과(공학사)
1987년 광운대학교 컴퓨터공학과(공학석사)
1993년 광운대학교 컴퓨터공학과(공학박사)
1994년 ~ 현재 배재대학교 컴퓨터공학과 교수
※관심분야 : 멀티미디어 문서정보처리, XML, Semantic Web, Ubiquitous Computing, USN, IoT