크롤러와 형태소 분석기를 활용한 웹상 개인정보 유출 판별 시스템

이형선[1] · 박재희[1] · 나철훈[2] · 정회경[1*]

[1]배재대학교 · [2]국립목포대학교 정보통신공학과

Crawlers and Morphological Analyzers Utilize to Identify Personal Information Leaks on the Web System

Hyeongseon Lee[1] · Park Jaehee[1] · Cheolhun Na[2] · Hoekyung Jung[1*]

[1]Paichai University · [2]Dept. of Information & comm. Eng. Mokpo National University

E-mail : prospace13@gmail.com, pjeahi0728@naver.com, chma@mokpo.ac.kr, hkjung@pcu.ac.kr

요 약

최근 개인정보 유출 문제가 대두됨에 따라 데이터 수집과 웹 문서 분류에 관한 연구들이 이루어지고 있다. 기존 시스템은 개인정보의 유무 여부만 판단하고 동명이인이나 사용자가 게시한 문서에 대한 분류는 이루어지지 않기 때문에 불필요한 데이터가 필터링 되지 않는 문제점이 있다.

본 논문에서는 이를 해결하기 위해 크롤러와 형태소 분석기를 활용하여 유출된 데이터의 유형이나 동음이의어를 식별할 수 있는 시스템을 제안한다. 사용자는 크롤러를 통해 웹상의 개인정보를 수집한다. 수집된 데이터는 형태소 분석기를 통해 분류한 후 유출된 데이터를 확인할 수 있다. 또한 시스템을 재사용 할 경우 정확도가 더 높은 결과를 얻을 수 있다. 이를 통해 사용자는 맞춤형 데이터를 제공 받을 수 있을 것으로 사료된다.

ABSTRACT

Recently, as the problem of personal information leakage has emerged, studies on data collection and web document classification have been made. The existing system judges only the existence of personal information, and there is a problem in that unnecessary data is not filtered because classification of documents published by the same name or user is not performed.

In this paper, we propose a system that can identify the types of data or homonyms using the crawler and morphological analyzer for solve the problem. The user collects personal information on the web through the crawler. The collected data can be classified through the morpheme analyzer, and then the leaked data can be confirmed. Also, if the system is reused, more accurate results can be obtained. It is expected that users will be provided with customized data.

키워드

Classification, Crawler, Morphological Analyzers, Privacy, Web Document

Ⅰ. 서 론

최근 개인정보 유출 문제가 대두됨에 따라 데이터 수집과 웹 문서 분류에 대한 연구들이 진행되고 있다[1,2]. 기존 시스템들은 SNS 내의 프라이버시 유출 사전 탐지, 개인정보 별 위험도 측정 방법 등을 통해 사용자에게 개인정보 유출 판별 서비스를 제공한다. 이에 반해 기존 시스템은 사전 예방에 중점을 두었기 때문에 이미 유출된 개인정보를 식별할 수 없고 개인정보 유포자에 대한자동 관리가 이루어지지 않은 문제점이 있다[3].

본 논문에서는 이를 해결하기 위해 크롤러와 형태소 분석기 기반 웹상 개인정보 유출 판별 시스템을 제안한다. 크롤러를 통해 사용자가 입력한 키워드를 포함하고 있는 웹 문서를 수집하고 형태소 분석기를 통해 문장에서 주어와 특정 정규식에 해당되는 텍스트 데이터를 필터링 한다. 또한 유출된 데이터의 링크나 유출 데이터 유형 등을 데이터베이스에 지속적으로 적재함으로써 정확도를 높이고 유출된 데이터와 유포자를 감시한다. 이를 통해 사용자에게 기존의 시스템보다 높은 정확도의 서비스를 제공할 수 있을 것으로 사료된다.

Ⅱ. 시스템 설계

본 장에서는 제안하는 시스템의 설계에 대해 다룬다. 그림 1은 시스템의 구조도이다.

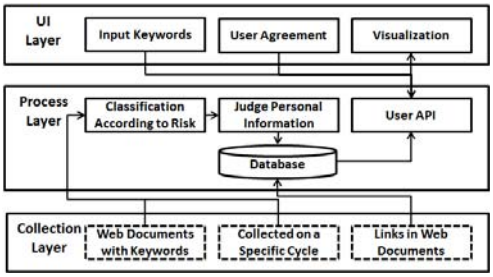

그림 1. 시스템 구조도

Collection Layer에서는 크롤러를 통해 사용자가 입력한 키워드 수집, 해당 문서의 링크를 데이터베이스에 적재하고 특정 주기마다 데이터를 수집할 수 있는 모듈로 구성된다. Process Layer는 Collection Layer에서 수집된 데이터를 통해 위험도에 따라 데이터를 분류하는 모듈과 개인정보 여부를 판단하여 데이터베이스에 적재하는 모듈 그리고 사용자와의 통신을 위한 User API로 구성되어 있다. UI Layer는 개인정보 수집 및 이용 동의, 키워드를 입력하는 페이지와 유출된 데이터를 보여주는 페이지 등의 UI로 이루어져 있다.

그림 2는 시스템의 흐름을 나타낸다. 어플리케이션이 실행되고 사용자가 개인정보 수집 및 이용 약관에 동의 했을 경우 사용자는 이름, 전화번호, 주소 등의 개인정보를 입력 한다. 입력한 데이터에 정규식을 적용하여 다양한 형태로 데이터를 변환한다. 변환 후 사용자가 입력한 키워드와 변환된 정보를 크롤러를 통해 수집한다. 수집된 데이터에서 솔루션 제공 알고리즘을 통해 사용자의 개인정보가 유출되었는지 여부를 판별하고 개인정보가 유출된 게시글의 링크 및 개인정보 유포자의 정보를 데이터베이스에 적재한다. 데이터베이스에 적재한 개인정보를 서버에서 분류한다. 분류된 데이터를 User API를 통해 추출한다. 추출한 데이터를 어플리케이션을 통해 사용자에게 제공한다.

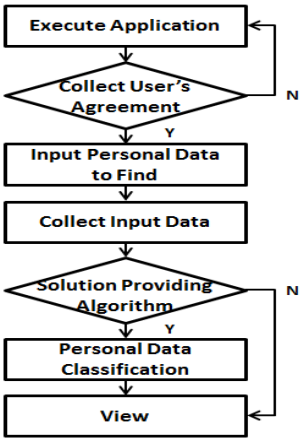

그림 2. 시스템 흐름도

Ⅲ. 결 론

최근 소셜 미디어나 SNS등의 서비스가 발달함에 따라 개인정보 유출문제가 대두되고 있다. 기존의 시스템들은 웹상에 유출된 개인정보를 식별할 수 없고 유포자에 대한 자동관리가 이루어지지 않는 문제점이 있었다.

본 논문에서는 이를 해결하기 위해 크롤러와 형태소 분석기 기반 웹상 개인정보 유출 판별 시스템을 제안하였다. 이를 통해 사용자에게 기존 시스템보다 높은 정확도의 서비스를 제공할 수 있을 것으로 사료되며 자동 관리 서비스를 제공할 수 있을 것이다.

향후 연구로는 본 논문에서 제안하는 시스템을 적용하고 효율성을 검증하기 위한 실험을 진행해야 할 것이다.

참고문헌

[1] M. H. Cheon, J. S. Choi, Y. T. Shin, "Measuring Method of Personal Information Leaking Risk Factor to Prevent leak of Personal Information in SNS" *Journal of the Korea Institute of Information Security & Cryptology*, vol. 23, no. 6, pp. 1199-1206, 2013.

[2] J. E. Park, M. S. Park, S. J. Kim, "The Reliability Evaluation of User Account on Facebook" *Journal of the Korea Institute of Information Security*, vol. 23, no. 6, pp. 1087-1101, 2013.

[3] H. T. Chae, S. J. Lee, "Security Policy Proposals Through PC Security Solution Log Analysis-Prevention Leakage of Personal Information." *Journal of the Korea Institute of Information Security*, vol. 24, no. 5, pp. 961-968, 2014.