Advanced Science and Technology Letters
Vol.87 (Art, Culture, Game, Graphics, Broadcasting and Digital Contents 2015), pp.42-46
http://dx.doi.org/10.14257/astl.2015.87.10

A Personal Information Retrieval System in a Web Environment

YoungDeok Seo, JunHyung Oh, JaeYoung Chang, Il-Min Kim

Department of Computer Engineering, Hansung University,
389 Samsun-Dong 3-Ga, Sungbuk-Gu, Seoul, South Korea
tera16@naver.com, ojunhy0@gmail.com, jychang@hansung.ac.kr, ikim@hansung.ac.kr

Abstract. Since we use internet every day, the internet privacy has become important. We need to find out what kinds of personal information is exposed to the internet and to eliminate the exposed information. By searching fragmentary clues using web search engines, a person can determine whether the personal information is exposed or not. This method has lots of problems. In this paper, we implemented a personal information retrieval system and proposed a method to remove one's private data from the Web easily.

Keywords: Internet privacy, Personal Information, Web Search Engine

1 Introduction

In 2014, three major credit card companies in South Korea were hacked and the 105 million personal data were stolen. This accident prompted the public concern for protecting private data. Some people may search the exposed personal information in Google with fragmentary keywords. However, since the results of the search engine are generally based on the accuracy and popularity of web pages, the searching capability focused on private data would be limited. Moreover, using a search engine directly for detecting the personal information could be dangerous because it would be susceptible to phaming or other phishing attacks.

Lots of researches have been conducted for protecting the private data which were exposed to web environment. Their main concerns were protecting large scale business data. The researches were very effective for reducing the possibility of data leaking and enhancing the sensitiveness of private data patterns. However, the researches on the private data management using search engines were scarcely ever conducted yet. In this paper, we proposed a new personal information retrieval system. The system can retrieve private data which are exposed to internet accesses, and then block the exposed data and remove them. Retrieval of personal information is implemented by ranking the web pages according to the exposure degree of personal information.

ISSN: 2287-1233 ASTL

2 Related Works

Some Researchers proposed the solution for protecting the personal information stored in standalone file systems or web environment such as SNS [1][2]. In [1], they developed the scanning and filtering solution for automatic diagnosing the exposure of personal information embedded in attached file or articles. With analyzed results, the related documents were prohibited to upload/download. Real-time monitoring sub-systems may send warning messages to the corresponding user and then erase the hazardous data. In [2], they proposed the solution for detecting the weakness and blocking the private data leakage in SNS. Once private data are exposed, the related SNS messages fast spread. By analyzing the characteristic of the fast spread speed and applying web crawling technology, they proposed a new method for monitoring private data in real-time. In addition, other researchers proposed various methods for retrieving private data and preventing private data leakages [3][4][5][6].

3 Private Data Retrieval System

3.1 Data Retrieval Process

In this paper, we designed and implemented a personal information retrieval system. This system searches the web pages including the user's personal information based on the user input. The searched web pages would be ranked in order of the exposure factor. Then, the user can check the personal information in each retrieved page. The system also provided the process to delete the personal information.

Fig. 1 shows the private data retrieval process, which was proposed in this paper. The original query denotes the basic personal information which is stored in the system. The extended query is composed of the basic information and the information gathered from well-known sites keeping private data. In order to reduce the retrieval time, we combined extended queries with commonly used combinations. The original searched results would be ranked by a searched engine. Then, we re-rank them in the order of the exposure factor explained in the next subsection.

3.2 Ranking Method

The objectives of a search engine are not only to find web pages containing just keywords, but also to select more accurate and popular web pages. The ranking method we proposed in this paper reflected the level of the private data exposure, and then re-ranked retrieved results pages in the order of on the exposure factor. We chose the Google's PageRank as a basic ranking tool. By considering the number and the quality of links to a page, Google's PageRank computes a rough estimate of how important the web site is. The underlying assumption is that more important websites are likely to receive more links from other websites. We selected critical personal data

and decided its weight by the survey. The survey result is shown as Table 1, in which the weight varies from 1 to 10.

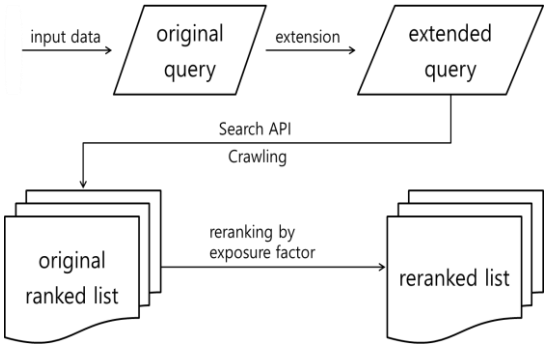


Fig. 1 Retrieval Process for Input Data

Table 2. Weights of Personal Information

Personal data	weight	Personal data (cont'd)	Weight (con'd)
SSN	7.2	Address	6.9
cellphone	6.1	employment	5.7
Phone	5.3	Birthday	4.9
account	6.1	Email	5.7

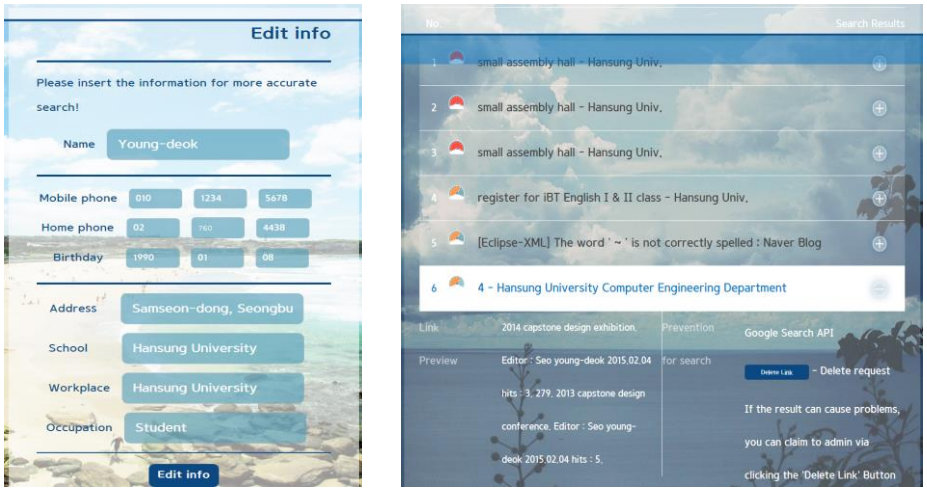
In this paper, we proposed a formula for evaluating personal information exposure factors. The equation is based on PageRank of Google and weights of personal information of Table 1 as follows:

$$E = PR \times \sum_{i=1}^{n} \left(\frac{1}{R}(1-R)^{KF_i} \times W_i\right) \quad (1)$$

In the equation above, E is the personal information exposure factor of a web page. The value of E is bigger or equal to 0. PR is the PageRank value of the page. KF_i ($1 \le i \le n$) is the keyword frequency which is the number of appearances of keyword i in the page. R is the correction factor, whose value is fixed to 0.5. Finally, W_i is the weight of a keyword i, which represents the importance of each personal information. The weights of Table 1 are applied in original queries. The half values of Table 1 are then applied in extended queries.

4 System Implementation

Java language was used for implementing the retrieval system. JSP(Java Server Page), HTML, CSS, JavaScript were also used to implement. MySQL database and CentOS 5.7 were used for the system.



(a) User data input **(b)** The search results

Fig. 2 An Example of the Personal Information Retrieval System

The user data input screen of the retrieval system is depicted in Fig. 2 (a). With the data, the system conducts web crawling and then displays the initial results. The system will re-rank the initial list by the exposure factor. The re-ranked list is depicted in Fig. 2 (b). If the user would click a link, the related snippet and the delete solution would be displayed.

5 Conclusion

In this paper, we designed and implemented a process searching web pages which contain the personal information. With this system, we can easily manage the exposed personal information in web environment. In the future work, we will further improve our work to be applied to business environment.

Acknowledgments. This research was supported by Basic Science research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant number: NRF-2011-0022445).

References

1. Whang, H., Kim, N.: Personal Information Protection System for Web Service, Journal of the Institute of Internet, Broadcasting and Communication, Vol. 11, No. 6, pp. 261-266 (2011)
2. Cho, H.:Design and Implementation of Personal Information exposure detection system in the SNS computing environment, MA thesis, SoongSil University (2013)
3. Cutrell, E., Dumais, S.T.: Searching to eliminate personal information management, Communications of the ACM, Vol. 49, No. 1, pp.58-64 (2006)
4. Dong, X., Halevy, A.: A platform for personal information management and integration, Proceedings of VLDB PhD Workshop (2005)
5. Krishnamurthy, B., Wills, C.E.: On the leakage of personally identifiable information via online social networks, Proceedings of the 2nd ACM workshop on Online social networks, pp7-12 (2009)
6. Irani, D., Webb, S., Pu, C., Li, K.: Modeling Unintended Personal Information Leakage from Multiple Online Social networks, IEEE Internet Computing, Vol 15, No. 3 (2011)

