International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 10 (2015) pp. 27165-27172
© Research India Publications
http://www.ripublication.com

Design and Implementation of a Personal Information Retrieval System in the Web

Young-deok Seo,Il-Min Kim,JaeYoung Chang

Department of Computer Engineering, Hansung University,
SungBuk-Gu, SamSunKyo-Ro 15Gil, Seoul,S.Korea
E-mail:tera16@naver.com, ikim@hansung.ac.kr, jychang@hansung.ac.kr

Abstract

Since internet became an important part of daily life, the need for protecting personal information has been increased. We need to find out what kind of personal information is exposed to the internet and to erase the exposed information from the internet. With the help of a popular web search engine such as Google, wecan determine whether personal information is exposed to the webor not. However, this method has a lot of limitations. In this paper, in order to overcome the limitations, we designed and implemented a new personal information retrieval system and proposed a new method to remove private data safely from theweb.

Keywords: Internet privacy, Web Search Engine, Personal Information

1 Introduction

The computer systems of major credit card companies in S. Korea were hacked and 105 million personal data were stolen in 2014. This accident prompted the public concern for protecting private data. According to the survey conducted by Trend Monitor at Sep. 2011, 94.3% of the 1,000 surveyees agreed that enterprises lacked the awareness of the importance of private data [1].The data which were exposed to the web could be easily accessed by anybody and could stay in the web for long time. For example, some employers wouldtry to search theweb in order to get the information of applicants.In such case, the web searching results might be critical to job appliers if the results aredifferentfrom what really they are.

Some people may search the exposed personal information in Google using fragmentary keywords. However, since the search engine results are generally based on the accuracy and popularity of web pages, the searching capability focused on private data would be limited. Moreover, typingprivate keywordsdirectly to a search engine for retrieving the personal information could be dangerous because it would be susceptible to pharming or other phishing attacks.

Lots of researches have been conducted for protecting the private data which were exposed to web environment. Some researchers proposed the solution for protecting the personal information stored in stand-alone file systems or web environment such as SNS [2][3]. In [2], they developed a scanning and filtering solution for automatic diagnosingthe exposure of personal information embedded in attached files or articles. With the use ofthe analysis results, the system prohibited the related documents from uploading/downloading. Real-time monitoring sub-systems may send warning messages to the corresponding users and then erase the hazardous data. In addition, other researchers proposed various methods for retrieving private data and preventing private data leakages [4][5][6][7]. Their main concerns were protecting large scale business data. Those researches were very effective for reducing the possibility of data leaking and enhancing the sensitiveness of private data patterns [12][13]. However, researches on managing the private datawithpopular 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, 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.If the exposure degree of a page is higher, the page is located on the top of the list. The system would not display the private data if the corresponding user does not want to. We conducted experiments to evaluate the accuracy of the ranking order. The experiment results showed that our search accuracy was greatly improved compared with the conventional web search engines.

2 Related Works

In South Korea, many leading privacy protection systems have been developedsuch as Siren24 and MarkAny. In order to prevent private information from being stolen, Siren24 sends a real-time alarm message whenever ID is authenticated in the web and is able to provide the history of the private information usage[8]. MarkAny prevents information leakage and the unauthorized usage of online content by managingauthorization[9].

In [3], they proposed the solution for detecting the weakness and blocking the private data leakage in SNS. Once private data were exposed, the related SNS messages spread rapidly. By analysing the characteristic of the fast spread speed and applying web crawling technology, they proposed a new method for monitoring private data in real-time. However, there were only few published researches onanalysing private data which were already exposed on the web.

3 System Compositions
3.1 Data Retrieval Process
In this paper, we designed and implemented a newpersonal information retrieval system for protecting private information. 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 could find out the personal information in each retrieved page. The system also provided methodsto delete the personal information from the web.

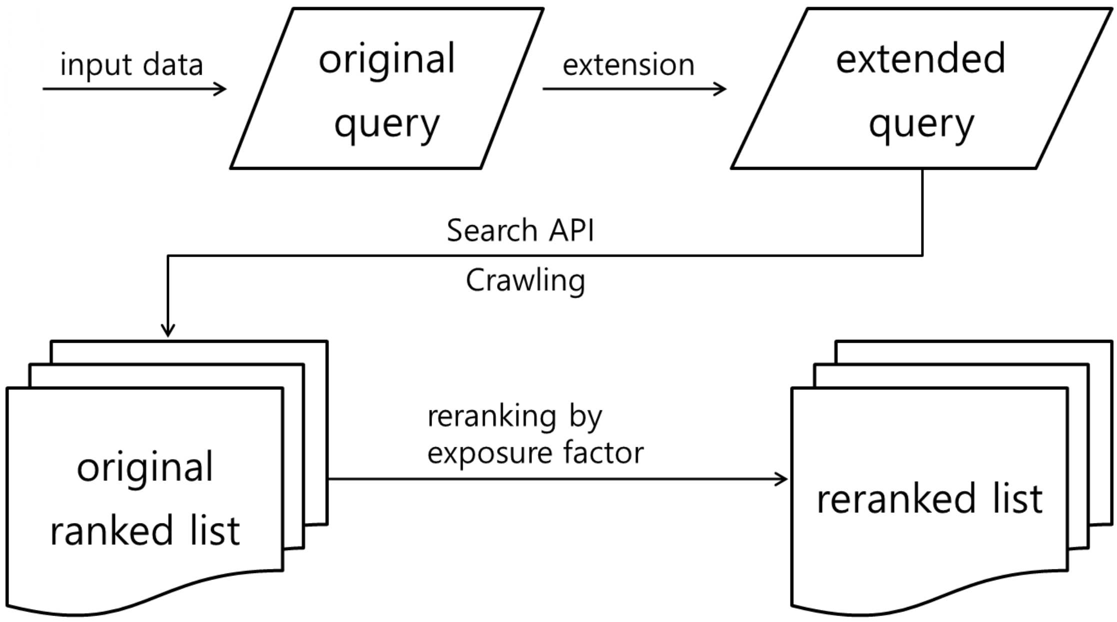

Fig. 1. Private data Retrieval Process

Fig. 1 depicts the private data retrieval process proposed in this paper. The original query denotes the basic personal information stored in the system. The extended query is composed of the basic information and the information gatheredfrom 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 such as Google. Then, we would re-rank them in order of the exposure degree. If the exposure value of the private data was 0, the corresponding link would be removed from the list.

3.2 Ranking Method
The objectives of a search engine are not only to find web pages containing the keywords, but also to select more accurate and popular web pages. The ranking method we proposed in this paper reflected the degree of the private data exposure, and then re-ranked the retrieved pages in order of 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[10] estimates the relative importanceof the web site. Computing a page ranking could be done as (1).

$$PR(A) = \frac{1-d}{N} + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)} \tag{1}$$

PR(A) is a PageRank value of document A, and T_i isa document pointing to A. *N* is the number of the total web documents, and $C(T_i)$ is the number of links which the document T_i has, and *d* is the damping factor. The damping factor of a page is the probability of moving toother pages because the page seems to be unsatisfied to the visitor. In general, the damping factor is fixed to 0.85.

In this paper, we developed an equation for evaluating the private data exposure factor (*EF*) for a page using a PageRank. The equation we developed is as follows.

$$EF(A) = PR(A) \times \sum_{i=1}^{n}\left(\frac{1}{R}(1-R)^{KF_i} \times W_i\right) \tag{2}$$

In equation (2),*PR* is the PageRank of the corresponding web page. KF_i ($1 \le i \le n$) is the keyword frequency which is the number of appearances of keyword *i* in the page A,and *R* is the correction factor. *R*is fixed to 0.5 in order to limit the variations of values.W_i is a weight of a keyword *i*, which represents the importance of each personal information (see Table 1).

We assigned the weights of personal information by the survey. The survey results were shown as Table 1 which were the result of our survey directly on 48 college students. The weight change was from 1 (the least important) to 10 (the most important). The values of Table 1 were applied in original queries. The half values of Table 1 were applied in extended queries.

Table 1. Weights of Personal Information

Personal data	Weight	Personal data	Weight
SSN	7.2	Home Address	6.9
cellphone	6.1	Work Address	5.7
Phone	5.3	Birthday	4.9
ID	6.1	Email	5.7
Name	5.4	Job	4.9
School	4.9		

The following is an example of the information retrieval process. The original query in the Table 2 refers to the keywords which are entered by the user. After the extended query is generated, the program collects web documents.

Table 2.Original Query & Extended Query

Original Query	{John Doe, Student, Hansung University, jkim@naver.com, 604-456-1243, *address* }
Extended Query	Original Query ∪{*id, nickname, etc* }

Table 3 shows the result of web documents collected byGoogle search engine. In this table, web pages ranked 1st, 2nd, 3rd, 8th, 9th and 10th were related to the user. On the other hand, pages ranked 4th, 5th, 6th and 7th were not related.

Table 3. List of Web Documents using Google

Ranking	Google Search Results
1	John Doe, Data Base – Hansung University
2	IT Project(MYPI) Report – Search Engine
3	Hansung University, Computer Engineering Dept. Homepage
4	Hansung University, alumni 1999, John Doe
5	Wikitree, Hansung Univ. apology for damaging a hand-written poster.
6	Hansung Univ. Korean Language Dept. Planning Dept.
7	Hansung Univ. Korean Language Dept. Planning Dept.
8	Hansung Univ. Foundation club – developing a app
9	Banner – Hansung University John Doe
10	Hansung Univ. Foundation club – developing a app

Table 4 shows re-ranked web pages in order of exposure factors. The exposure factors of Table 4 were computed by equation (2). Pages ranked 1st ~9th of the Table 4 were related the user. Especially, top-ranked page included private data such as user's name, school, and cellphone numbers. However, the page ranked 10th was not related to the user,even though the page includes user's name (a personwiththe samename).

Table 4. Re-ranked List and Exposure Factor

Ranking	Web Document	Exposure Factor
1	Hansung University auditorium lease – John Doe	167
2	Hansung University auditorium lease – John Doe	164
3	Hansung University auditorium lease – John Doe	164
4	Hansung Univ. English class – class register guide	73
5	Terabyte blog – Eclipse plug-in error debugging	67
6	Developer Email – xxx6@naver.com – Midamore app	20
7	Developer Email – xxx6@naver.com – Midamore app	18
8	Hansung Univ. Computer Eng. Dept. homepage	18
9	Hansung Univ. successful candidate list for edu. Class	15
10	List of 4776 pro-Japan collaborators released	10

4Implementation andPerformance

We mainly used Java language for implementing the personal data retrieval system. JSP(Java Server Page), HTML, CSS, JavaScript were used for the web implementation. MySQL database and CentOS 5.7 were used for the platform.The implementation details were published in [14].In this research, we used DCG (Discounted Cumulative Gain) [11] as a measure for comparing the search results. DCG is commonly used as a tool to measure the efficiency of search engines. DCG assigns a score to the search results depending on the relationship between the search keywords and the search results. In DCG,there are CG(Cumulative Gain) and nDCG (Normalized Discounted Cumulative Gain) methods. The CG method quantifies the relevance of search results andthen the computes the closeness. The nDCG normalizes the closeness from 0 to 1. CG computes closeness as the sum of all of numerical scores which reflect the relevance of search results regardless of the order. If the number of search results is *p*, CG uses the following equation.

$$CG_p = \sum_{i=1}^{p} rel_i \tag{3}$$

In the equation (3),rel_iindicates the relevance value of *i*-th result in the search results. DCG considers that highly relevant results appearing lower in the search results should be penalized as the graded relevance value. DCG is defined as follows.

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{4}$$

In general, the value computed by equation (4) is not used as it is. Instead, nDCG, which is a normalizedvalue of DCGinto (0, 1), is generally used. The equation for the normalization is as follows.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{5}$$

*IDCGp*would be same as DCGunder the assumption that the search results are correctly ranked by the relevance.Therefore, the more nDCG close to 1, the better the search result is. In this paper, we used nDCG for evaluating the search results. We assigned rel_i from 1 to 5 according to the degree of relationship between the search keywords and results. With no relationship, we assigned 0 to the value.

The performance results of the proposed system is depicted in Fig.2. We experimented seven users and the experiment results are T1 ~ T7 in Fig. 2. The rightest bar ‘Ave’ in Fig.2depicts the average. In Fig.2, *Original List* indicates the performance of searchingpersonal information using Google search engine. *Re-ranked List* indicates the performance of searching personal information using the system which was proposed in this paper.As shown in Fig.2, the performance of the proposed system was better than that of Google search engine in all cases. More specifically, the searching accuracy was improved from 19.81% to 67.79%. We found out the searching accuracy was improved 39.47% on average compared with Google.

5 Conclusion

In this research, we designed and implemented a new information retrieval system, which could search personal information more accurately and efficiently. We also included the performance result of the system in this paper. The system we proposed showed better performance than existing systems.However, since the personal information retrieval system is currently under development, the system still has the following problems which require improvements. First, in order to trace personal information, the system requires to sign up and register. Second, the weights of the personal information are fixed and do not allow change in the retrieval system. This means that every user has same weights of her personal information. In the future, we need to allow more personalized searches by adding a function to control the private weights.

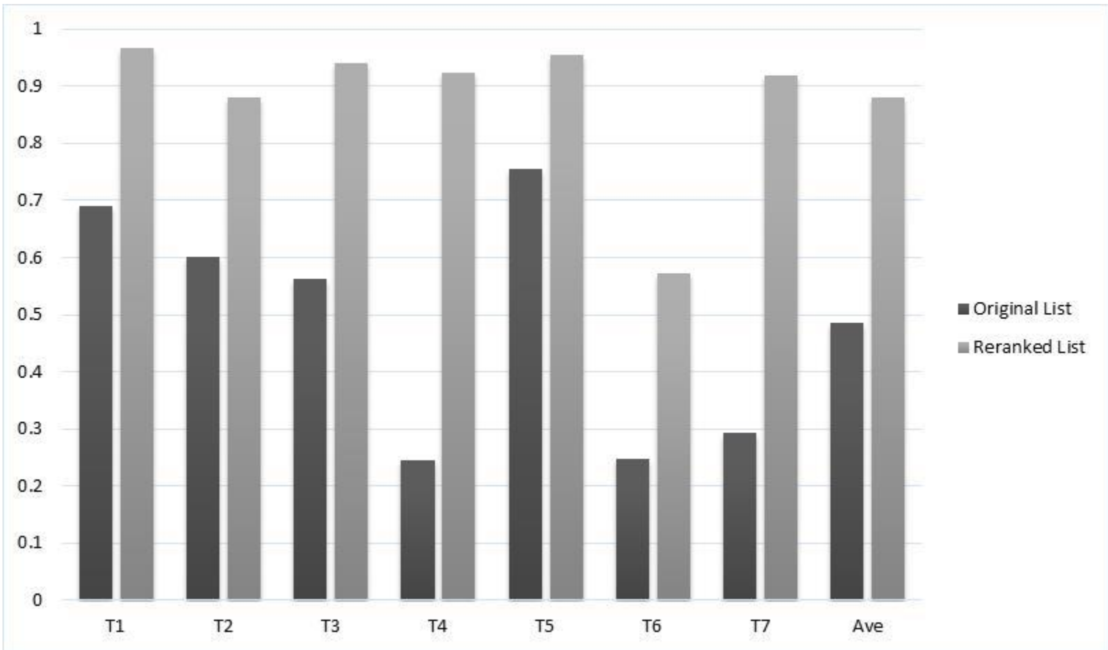

Fig. 2. Empirical results using nDCG

Acknowledgments

This research was financially supported by Hansung University.

References

[1] http://www.trendmonitor.co.kr/

[2] Whang, H., Kim, N., 2011,"Personal Information Protection System for Web Service", Journal of the Institute of Internet, Broadcasting and Communication, Vol. 11, No. 6, pp. 261-266.

[3] Cho, H., 2013, "Design and Implementation of Personal Information exposure detection system in the SNS computing environment", MA thesis, SoongSil University.

[4] Cutrell, E., Dumais, S.T., 2006,"Searching to eliminate personal information management", Communications of the ACM, Vol. 49, No. 1, pp.58-64.

[5] Dong, X., Halevy, A., 2005, A platform for personal information management and integration, Proceedings of VLDB Ph.D. Workshop.

[6] Krishnamurthy, B., Wills, C.E., 2009,"On the leakage of personally identifiable information via online social networks", Proceedings of the 2nd ACM workshop on Online social networks, pp7-12.

[7] Irani, D., Webb, S., Pu, C., Li, K., 2011,"Modeling Unintended Personal Information Leakage from Multiple Online Social networks", IEEE Internet Computing, Vol 15, No. 3.

[8] http://www.siren24.com/

[9] http://www.markany.com/

[10] http://en.wikipedia.org/wiki/PageRank/

[11] Kim, H., Chang, J., 2011, "Discovering News Keyword Associations Using Association Rule Mining", Journal of The Institute of Internet, Broadcasting and Communication Vol 11. No. 6, pp.63-71.

[12] Takahashi, K., Liu, Z., Sakurai, K., Amamlya, M., 2007, "A Framework for User Privacy Protection Using Trusted Programs", International Journal of Security and Its Applications, Vol. 1, No2, pp. 59-70.

[13] Parra-Arnau, J., Rebollo-Monedero, D., Forne, J., 2012, "A Privacy-Protecting Architecture for Recommendation System via the Suppression of Ratings", International Journal of Security and Its Applications, Vol.6, No.2, pp.61-80.

[14] YoungDeok Seo, JunHyung Oh, JaeYoung Chang, Il-Min Kim, A Personal Information Retrieval System in a Web Environment, Advanced Science and Technology Letters Vol.87 (Art, Culture, Game, Graphs, Broadcasting and Digital Contents 2015), pp.42-46