The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 15, No. 4, pp.145-151, Aug. 31, 2015. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2015.15.4.145
JIIBC 2015-4-18

웹 환경에서의 개인정보 검색기법

A Retrieval Technique of Personal Information in a Web Environment

서영덕*, 장재영**

Young-Duk Seo*, Jae-Young Chang**

요 약 최근 개인정보 보호에 대한 관심이 높아지면서 웹 환경에 노출된 개인정보를 효율적으로 검색할 수 있는 시스템의 필요성이 증대되었다. 현재 웹 환경에 노출된 개인정보는 자신이 알고 있는 단편적인 단서를 이용한 검색을 통해 노출여부를 판단하고 있다. 그러나 이 방법은 검색결과의 우선순위가 개인정보의 노출도와 관계가 없어 효율적이지 못하다. 본 논문에서는 사용자 입력을 근거로 웹 환경에 노출된 사용자의 개인정보를 효율적으로 검색하고 삭제할 수 있도록 지원하는 프로세스를 제안한다. 또한 기존 검색 방법과의 비교를 통하여 검색성능의 향상 정도를 평가한다.

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. However, it is not efficient to search the personal information using only fragmentary clues in web search engines because the ranking results are not relevant to the exposure degree of personal information. In this paper, we introduced a personal information retrieval system and proposed a process to remove private data from the web easily. We also compared our proposed method with previous methods by evaluating the search performance.

Key Words : Internet privacy, Personal Information, Web Search Engine, Ranking

Ⅰ. 서 론

2014년 1월 대형 카드 3사에서 약 1억 580만 건의 개인정보(personal information)가 유출되면서 개인정보 보호에 대한 관심이 폭발적으로 증가하였다. 개인정보 유출에 대한 우려는 이미 이전부터 지속적으로 제기되고 있었다. 2011년 9월 시장조사 전문기업 Trend Monitor가 인터넷을 이용하는 성인 남녀 1000명을 대상으로 실시한 개인정보 유출 관련 인식 조사에 따르면 <개인정보의 중요성에 대한 기업들의 인식이 부족하다>에 대한 항목 동의율은 94.3%, <사이트 가입 시 기업에서 요구하는 개인정보 수집은 최소화해야 한다>에 대한 항목 동의율은 93.3%로 응답자 특성별 큰 차이 없이 공감대를 형성하고 있는 것을 알 수 있다[1].

웹 환경에 공개적으로 노출된 개인정보는 누구나 검색하여 찾을 수 있으며, 당사자가 직접 관리하지 않는다

*준회원, 한성대학교 컴퓨터공학과
**정회원, 한성대학교 컴퓨터공학과(교신저자)
접수일자 2015년 7월 14일, 수정완료 2015년 8월 7일
게재확정일자 2015년 8월 7일

Received: 14 July, 2015 / Revised: 2 August, 2015 /
Accepted: 7 August, 2015
**Corresponding Author: jychang@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

면 영구적으로 남게 될 가능성이 있다. 일부 대학과 고용업체에서는 지원자에 대한 정보를 얻기 위해 웹을 직접 검색하기도 한다. 검색된 정보는 본인의 의도와는 다른 방식으로 해석될 가능성이 있기 때문에 어떤 정보가 인터넷에 공개되었는지 파악하고 그 정보가 타인들에게 주는 인상을 확인하는 것은 필수적이다.

현재 사람들은 자신과 관련된 단편적인 키워드를 무작위로 Google, Naver 등의 대형 웹 엔진에 검색하는 방법으로 웹 환경에 노출된 자신의 개인정보를 찾아 관리하고 있다. 하지만 일반적인 웹 엔진은 검색한 키워드를 중심으로 하여 색인 페이지(indexed page)의 정확도(accuracy)와 인기도(popularity)를 기반으로 한 알고리즘을 사용하기 때문에 개인정보의 노출현황과 그 정도를 파악하기에는 검색 능률이 떨어진다. 또한, 개인이 직접 개인정보 키워드를 웹 엔진에 입력하여 검색하는 경우에는 파밍(pharming) 등의 2차적인 보안 피해가 일어날 수 있는 위험성도 존재한다.

개인정보 유출에 대한 관심이 높아지면서 개인 컴퓨터 내의 파일이나 웹 환경 등에 노출된 개인정보를 보호하기 위한 많은 연구가 진행되어 왔다. 대표적으로 [2]와 [3]에서는 내부 시스템 또는 웹 크롤링(crawling)을 통해 웹 환경에 노출된 개인정보를 분석하여 해당 문서의 다운로드/업로드를 금지시키거나 실시간 모니터링을 통해 경고, 삭제를 유도하는 방법을 제안하였다. 이외에도 SNS를 비롯한 다양한 환경에서 개인정보의 유출을 방지하는 기법들이 제안되었다.[4][5][6][7][12] 이러한 연구들은 주로 기업체 등의 대형 데이터를 대상으로 개인정보가 유출될 가능성을 사전에 차단하고 개인정보 관련 패턴에 대한 민감성을 증가시키는 데 큰 기여를 하였다. 그러나 검색 엔진을 통해 개인정보를 관리할 수 있도록 하는 시스템에 대한 연구는 미진하다.

본 논문에서는 웹 환경에 노출된 개인정보를 검색하여 이를 차단, 삭제하도록 지원하는 개인정보 검색 시스템을 제안한다. 제안된 시스템에서는 사용자 입력을 근거로 웹 환경에 남아있는 개인정보를 노출정도에 따라 랭킹(ranking)하여 보여주도록 구성하였다. 랭킹은 기본적으로 Google의 PageRank를 기반으로 하였으며, 개인정보의 유형에 따른 노출정도를 파악한 후 노출도가 높은 페이지를 상위에 랭크되도록 설계하였다. 또한 검색된 개인정보가 노출되기를 원치 않을 경우 이를 검색엔진의 결과에서 배제되도록 요청할 수 있는 환경을 개발하였다. 랭킹의 정확도를 평가하기 위해 실험을 실시하였으며 실험결과 기존의 웹 검색엔진에 비해 개인정보의 검색 정확도가 크게 향상된 것을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 소개하고, 3장에서는 개인정보 검색을 위한 과정과 랭킹 기법을 소개한다. 4장에서는 개인정보 검색시스템을 구현한 내용을 기술하며 5장에서는 시스템의 성능을 평가하고 마지막으로 6장에서는 결론을 맺는다.

II. 관련연구

현재 국내의 대표적인 개인정보 보호시스템으로는 사이렌24, 마크애니 등이 있다. 사이렌24는 개인정보가 타인에 의해 도용되는 것을 방지하기 위하여 개인정보 인증 발생 시 실시간 알람을 발송하고 정보 이용 내역을 조회할 수 있는 기능 등을 구비하고 있다[8]. 마크애니는 정규식(regular expression) 및 키워드 정보를 활용하여 문서 내 개인정보 필터링을 통해 정보 유출을 방지하고 있다[9].

관련 연구로서 [2]에서는 홈페이지 상의 게시판 및 첨부파일을 대상으로 자동화된 개인정보 노출 진단을 위한 스캐닝 및 필터링 솔루션을 개발하였다. 이 솔루션은 파일 캐싱(caching) 및 압축 기법을 활용한 웹 가속 기능을 통해 신속하게 기업 내부의 보안 취약점을 발견하고 정보 유출을 차단한다. 또한 [3]에서는 개인정보 노출 시 급속도로 확산되는 SNS의 특성을 분석하여 웹 크롤링 기술과 개인정보 패턴 분석 및 오탐(false positive) 제거 프로세스를 적용하였다. 이를 통해 해당 정보의 정확성을 분석하고 개인정보를 실시간 모니터링 할 수 있는 방법을 제안하였다. 그러나 현재 웹 환경에 이미 노출된 데이터들을 대상으로 개인화된 검색을 통하여 개인정보를 노출정도에 따라 분석하여 관리하는 시스템에 대한 공개된 연구는 거의 없다.

III. 개인정보 검색기법

1. 검색 프로세스

본 논문에서 제안하는 검색 시스템에서는 입력된 정보를 이용하여 웹 환경에 노출된 개인정보를 포함한 웹

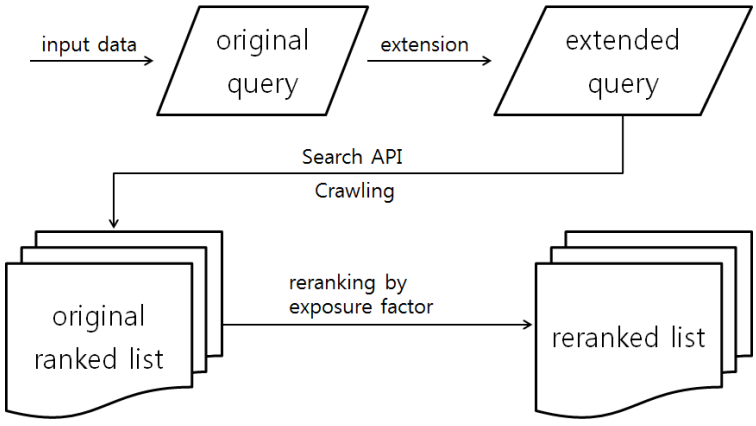

그림 1. 개인정보 검색 프로세스
Fig. 1. Retrieval Process of Personal Information

문서를 노출도순으로 랭킹하여 보여준다. 또한 사용자는 랭킹된 웹 문서들에 대한 간략한 정보와 함께 해당 정보를 삭제할 수 있는 솔루션을 제공받게 된다.

그림 1은 본 논문에서 제안하는 개인정보 검색 과정을 보여준다. 이 그림에서 original query는 사용자가 초기에 입력한 간단한 개인정보들이다. 개인정보 검색은 이 정보만으로 수행할 수 있으나, 보다 정확한 검색을 위해서 이 정보를 이용하여 개인정보가 노출된 잘 알려진 사이트로부터 추가적인 정보들을 수집한다. 그리고 이들 정보를 결합하여 extended query를 구성한다. 이렇게 구성된 개인정보 리스트를 여러 가지 조합으로 구성한 후 query들을 생성하고 Google Search API를 통하여 검색결과를 수집한다. 여기서 수집된 결과가 그림 1에서 original ranked list이다. 마지막으로 이 정보를 기반으로 각 웹 문서에 대해 개인정보 노출도(exposure factor)를 계산하여 노출도가 높은 순으로 재정렬(reranking)한다. 재정렬된 리스트 중 개인정보 노출도가 0인 웹 문서는 리스트에서 제외한다.

2. 랭킹 기법

검색 엔진에 검색어를 입력하는 목적은 검색어가 포함된 문서만을 찾기 위한 것이 아니라 그 중 정확도나 인기도가 가장 높은 문서를 찾는 것이다. 본 논문에서는 각 문서의 개인정보 노출 정도를 평가하기 위해 효율적인 랭킹 알고리즘으로 알려져 있는 Google의 PageRank 기법을 일부 활용하였다.

PageRank[10]는 하이퍼링크(hyperlink) 구조를 가지는 문서가 있을 때 대상 페이지에서 다른 페이지로 나가는 링크(outlink)와 다른 페이지에서 대상 페이지로 들어오는 링크(inlink)를 통하여 상대적 중요도를 판단하는 방식이다. 페이지랭크를 계산하는 방법은 식 (1)과 같다.

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

$PR(A)$는 문서 A의 PageRank 값이고, T_i는 그 문서를 가리키는 다른 문서, N은 웹 문서의 총 개수, $C(T_i)$는 T_i 문서가 가지고 있는 링크의 총 개수, d는 damping factor이다. damping factor란 해당 페이지에 만족하지 못하고 다른 페이지로 가는 링크를 클릭할 확률이며 일반적으로 0.85로 값을 설정한다.

본 논문에서는 PageRank 값을 활용한 개인정보 노출도를 평가하는 수식을 개발하였으며 다음의 식 (2)와 같다.

$$E(A) = PR(A) \times KC(A) \times \sum_{i=1}^{n}(\frac{1}{R}(1-R)^{KF_i} \times W_i) \quad (2)$$

여기서 $E(A)$는 문서 A의 노출도이고, $PR(A)\,PR(A)$는 해당 웹 문서의 PageRank 값이며, $KC(A)$는 키워드 결합도(keyword coupling)이다. 키워드 결합도는 웹 문서 내에서 포함하고 있는 서로 다른 개인정보의 종류에 기반하여 값이 결정된다. KF_i는 키워드 빈도수(keyword frequency)이고, R은 보정 계수(correction

factor)로 키워드 빈도수에 의한 값 변동 폭을 최대 2배로 제한하기 위하여 값을 0.5로 고정하였다. W_i는 해당 키워드의 가중치로 개인정보의 중요도에 따라 표 1과 같이 부여하였다.

표 1. 개인정보 중요도에 따른 가중치 값
Table 1. Weights of Personal Information

개인정보	가중치
주민등록번호	7.2
핸드폰번호	6.1
집전화번호	5.3
ID	6.1
이메일	5.7
이름	5.4
집주소	6.9
직장	5.7
생년월일	4.9
학교	4.9
직업	4.9

표 1에서 가중치는 각 개인정보의 중요도를 나타내며 최소값을 1, 최대값을 10으로 설정하였다. 이 값은 일반 대학생 48명을 대상으로 직접 조사한 개인정보별 중요도의 평균값이다. original query의 경우 위의 가중치 값을 직접 반영하며, extended query에서 original query를 제외한 나머지 즉, 새로이 확장된 query의 경우에는 해당 개인정보 가중치의 반을 부여하였다.

다음은 검색과정의 예를 보여준다. 표 2의 original query는 사용자가 초기에 입력한 검색어를 의미한다. 여기서 확장 과정을 거쳐 extended query가 생성되면 프로그램에서는 해당 query를 조합하여 웹 문서를 수집한다.

표 2. 원래의 쿼리와 확장된 쿼리
Table 2. Original Query and Extended Query

Original Query	{홍길동, 학생, 한성대 cxxxxi@naver.com, 010-5126-xxxx, 02-760-xxxx, 서울시 성북구 삼선동}
Extended Query	Original Query ∪ {txxxx6(id), xxx(nickname), 테라xxx(nickname)}

표 3은 Google 검색엔진의 웹 문서 수집 결과를 보여준다. 수집결과에서 No 1, 2, 3, 8, 9, 10은 사용자와 연관된 웹 문서이지만, No 4, 5, 6, 7은 실제 관련이 없는 웹 문서이다.

표 3. Google을 이용한 검색결과 리스트
Table 3. List of Web Documents using Google

No.	Google 검색결과
1	홍길동_데이터베이스 - 한성대학교
2	설계프로젝트(MYPI) 결과보고서 - 검색 엔진
3	한성대학교 컴퓨터공학과 홈페이지
4	한성대학교 총동문회 1999, 홍길동
5	wikitree| 한성대 대자보 훼손 영상 사과문
6	한성대학교 한국어문학부 기획부
7	한성대학교 한국어문학부 기획부
8	한성대 창업동아리, '응답하라 통학버스'앱 개발
9	지존현수막 - 한성대(컴퓨터공학과) 홍길동
10	한성대 창업동아리, '응답하라 통학버스'앱 개발

표 4. 재랭킹된 웹 문서 검색결과 리스트와 노출도
Table 4. Reranked List and Exposure

No	Web Document	노출도
1	한성대 소강당 대여 - 오케스트라 동아리 홍길동	167
2	한성대 소강당 대여 - 오케스트라 동아리 홍길동	164
3	한성대 소강당 대여 - 오케스트라 동아리 홍길동	164
4	한성대 영어 강좌 수강신청 관련 강의실 안내	73
5	테라바이트 블로그 - Eclipse plup-in error 디버깅	67
6	Developer Email : txxxx6@naver.com - 미다모아 앱	20
7	Developer Email : txxxx6@naver.com - 미다모아 앱	18
8	한성대학교 컴퓨터공학과 홈페이지	18
9	2010학년도 교직이수예정자 선발자 명단 - 한성대학교	15
10	한반도 시나리오 | 친일인사 4776명 공개	10

표 4는 본 논문에서 구현한 프로그램을 통하여 수집한 웹 문서를 노출도 기준으로 재랭킹한 결과이다. 이 표에서 노출도는 식 (2)로부터 계산된 결과이다. 본 결과의 No. 1 ~ 9는 사용자와 관련된 데이터이다. No. 1의 경우, 사용자의 이름, 학교, 휴대폰번호 정보가 다량 노출되어 있었다. 반면 No. 10의 경우, 사용자의 이름이 발견되었으나 사용자와는 관계없는 웹 문서로 판명되었다.

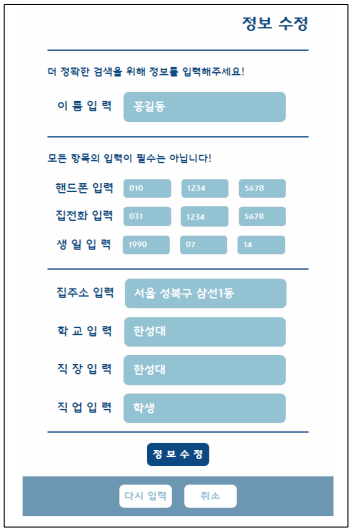

그림 2. 초기 사용자 데이터 입력
Fig. 2. Initial Personal Data Input

그림 3. 개인정보 검색결과 화면
Fig. 3. A Search Result Screen of Personal Information

Ⅳ. 시스템 구현

본 논문에서 개발한 개인정보 검색시스템의 개발환경은 다음과 같다. 개발 언어로는 Java를 주로 사용하였으며, JSP(Java Server Page), HTML CSS, JavaScript로 웹을 구성하였다. DB는 mysql을 사용하며, 서버는 CentOS 5.7을 사용하였다.

그림 2는 사용자 데이터를 입력하는 화면이다. 입력된 사용자 데이터를 기반으로 웹 크롤링을 진행하여 초기 웹 문서 리스트를 구성한다.

구성된 리스트를 노출도에 따라 재랭킹한 리스트가 그림 3이다. 리스트를 클릭하면 해당 웹 문서의 간략한 내용(snippet)과 함께 삭제 솔루션을 제공한다.

Ⅴ. 성능 평가

1. 성능평가 기준

본 논문에서 검색 결과를 비교하기 위한 측정치로서 DCG(Discounted Cumulative Gain)를 활용하였다[11]. DCG는 검색엔진의 효율성을 측정하는 도구로 많이 쓰이는 기법 중 하나이다. DCG는 검색결과로 나온 용어에 대해 검색어와의 실제 관련도에 따라 점수를 부여하는 방식을 사용한다. DCG에는 검색 결과의 순서(랭킹)보다 결과들의 관련도만을 정량화하여 계산하는 CG(Cumulative Gain) 방식이 있고, 검색결과의 랭킹을 중요한 요소로 활용하는 정규화된 측정방식으로 nDCG(Normalized Discounted Cumulative Gain)가 있다. 우선 CG는 검색결과의 순서에 관계없이 각 결과 단어의 관련정도를 점수화 하여 그 값을 모두 합한 수치로 평가한다. 즉, 검색결과가 p개일 경우 CG는 다음의 수식으로 계산한다.

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

여기서 rel_i는 검색 결과에서 i번째 위치한 단어의 관련도를 나타낸다. 반면에 DCG는 관련도가 높은 단어가 검색 결과에 우선적으로 랭킹된다는 사실에 입각해서 그렇지 않을 경우 감점을 부과하는 방식을 사용한다. 다라서 DCG는 다음의 수식으로 계산한다.

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

일반적으로 DCG는 위의 수식 그대로 사용하지 않고 0부터 1까지의 정규화된 수치인 nDCG로 표현하는데 수식은 다음과 같다.

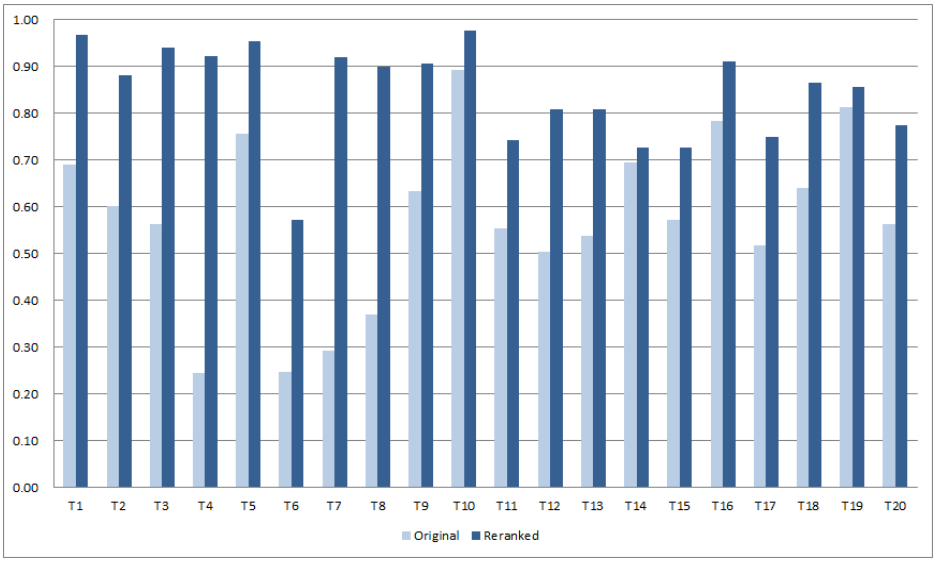

그림 4. nDCG를 이용한 성능평가결과
Fig. 4. Empirical results using nDCG

$$nDCG_p = \frac{DCG_p}{IDCG_p} \quad (5)$$

$IDCG_p$는 관련도 값에 따라 정확하게 정렬되었다는 가정하의 DCG값을 의미한다. 따라서 nDCG는 1에 가까울수록 좋은 검색 결과를 나타낸다. 본 논문에서는 nDCG를 검색 결과의 평가에 활용하였다. 관련도 평가를 위한 rel_i는 실제 개인정보 포함 여부 및 중요성에 따라 1에서 5사이의 값을 부여하였으며, 연관성이 없을 경우 0을 부여하였다.

2. 성능평가 결과

본 논문에서 구현한 개인정보 검색시스템의 성능평가 결과는 그림 4와 같다. 검색 성능은 사용자 20명에 대하여 테스트를 진행하였으며 그 결과는 각각 T1~T20과 같다.

이 그림에서 'Original'는 Google 검색결과에 대한 개인정보 검색 성능 나타내며, 'Reranked'는 본 논문에서 제안된 개인정보 검색 기법을 이용하여 랭킹한 리스트의 성능을 나타낸다. 이 그림에서 보는 바와 같이 실험 대상의 모든 경우에 대해서 본 논문이 제안한 개인정보 검색 기법이 기존의 웹 검색 엔진에서의 검색에 비해 우수한 성능을 보였다. 개인정보 관련 검색 정확도 향상률은 적은 경우 4.43%부터 많은 경우 67.79%로 측정되었으며 평균적으로 27.21% 향상되었음을 확인할 수 있었다.

Ⅵ. 결 론

본 논문에서는 개인정보가 포함된 웹 문서를 효율적으로 검색하는 기법을 제안하였으며 이를 구현한 결과를 소개하였다. 또한 성능 평가 결과 기존의 웹 검색에 비해 우수한 검색성능을 보였다.

본 논문은 개인정보 검색시스템 개발의 초기 모델로 다음과 같이 개선해야할 문제점이 있다. 첫째는 회원가입 시 입력되는 개인정보가 또 다른 정보유출의 원인이 될 수 있다는 점이다. 이는 기업 또는 보안회사와의 연계를 통하여 해결할 수 있을 것으로 판단된다. 둘째는 개인정보의 중요도가 고정되어있다는 점이다. 이는 향후 시스템 개선 시 개인정보 가중치를 개인이 조정할 수 있도록 하는 기능을 추가하여 보다 개인화된 검색을 보장하도록 하는 방법을 고려하고 있다.

References

[1] http://www.trendmonitor.co.kr/
[2] H. Whang and N. Kim, Personal Information

Protection System for Web Service, Journal of the Institute of Internet, Broadcasting and Communication, Vol. 11, No. 6, pp. 261-266, 2011.
[3] H. Cho. Design and implementation of personal information exposure detection system in the SNS computing environment. MAthesis, Soongsil Univ., 2013.
[4] E. Cutrell and S. T. Dumais, Searching to eliminate personal information management, Communications of the ACM, Vol. 49, No. 1, pp. 58-64, 2006.
[5] X. Dong and A Halevy, A platform for personal information management and integration, Proceedings of VLDB 2005 Ph.D. Workshop, 2005.
[6] B. Krishnamurthy, C. E. Wills, On the leakage of personally identifiable information via online social networks, '09 Proceedings of the 2nd ACM workshop on Online social networks, pp. 7-12. 2009.
[7] D. Irani, S. Webb, C. Pu, K. Li, Modeling Unintended Personal Information Leakage from Multiple Online Social Networks, IEEE Internet Computing, Vol. 15, No. 3, 2011.
[8] http://www.siren24.com
[9] http://www.markany.com
[10] http://en.wikipedia.org/wiki/PageRank
[11] H. Kim and J. Chang, Discovering News Keyword Associations Using Association Rule Mining, The Institute of Internet, Broadcasting and Communication Vol. 11. No. No. 6, pp. 63-71, 2011.
[12] H. Hwang, and N. Kim, Personal Information Protection System for Web Service, The Journal of The Institute of Internet, Broadcasting and Communication(JIIBC), Vol. 11, No. 6, Dec., 2011.
[13] J. Shim, H. C. Lee, "The Development of Automatic Ontology Generation System Using Extended Search Keywords" Journal of the Korea Academia-Industrial cooperation Society, Vol. 11, no. 6, 2009.

저자 소개

서 영 덕(준회원)

- 2015년 : 한성대학교 컴퓨터공학과 (공학사)
- 2015년 ~ 현재 : 한성대학교 컴퓨터 공학과 조교

<관심분야 : 정보검색, 데이터마이닝>

장 재 영(정회원)

- 1992년 : 서울대학교 계산통계학과 (이학사)
- 1994년 : 서울대학교 계산통계학과 (이학석사)
- 1999년 : 서울대학교 계산통계학과 (이학박사)
- 2000년 ~ 현재 : 한성대학교 컴퓨터 공학과 교수

<관심분야 : 데이터베이스, 정보검색, 데이터마이닝>

※ 본 연구는 한성대학교 교내학술연구비 지원과제임.