http://dx.doi.org/10.5392/JKCA.2011.11.9.009

웹 자원 아카이빙을 위한 웹 크롤러 연구 개발

Development of Web Crawler for Archiving Web Resources

김광영, 이원구, 이민호, 윤화묵, 신성호
한국과학기술정보연구원 정보기술연구실

Kwang-Young Kim(kykim@kisti.re.kr), Won-Goo Lee(wglee@kisti.re.kr),
Min-Ho Lee(cokeman@kisti.re.kr), Hwa-Mook Yoon(hmyoon@kisti.re.kr),
Sung-Ho Shin(maximus74@kisti.re.kr)

요약

웹 자원은 아직 수집, 보존, 활용에 대한 방안이 없어서 일정 기간의 서비스가 끝나면 사라져 버리는 문제점이 있다. 이런 웹 자원들은 중요성에 관계없이 주기적 또는 비주기적으로 갱신되거나 소멸된다. 따라서 웹 자원을 수집하고 보존하기위한 웹 아카이빙 시스템이 요구되고 있다. 이러한 웹 자원들을 주기적으로 수집하기 위해서는 웹 아카이빙 전용 크롤러의 개발이 필요하다. 따라서 본 연구에서는 웹 자원의 아카이빙 수집을 위해서 사용되는 기존의 웹 크롤러의 장단점을 분석하고 이것을 이용하여 웹 정보자원을 수집하기 위한 가장 적합한 수집 도구 시스템을 연구하고 개발하였다.

■ 중심어 : | 웹 자원 아카이빙 | 웹 크롤러 | 웹 스냅샷 로봇 | 아카이빙 시스템 | 영구 보존 |

Abstract

There are no way of collection, preservation and utilization for web resources after the service is terminated and is gone. However, these Web resources, regardless of the importance of periodically or aperiodically updated or have been destroyed. Therefore, to collect and preserve Web resources Web archive is being emphasized. Web resources collected periodically in order to develop Web archiving crawlers only was required. In this study, from the collection of Web resources to be used for archiving existing web crawlers to analyze the strengths and weaknesses. We have developed web archiving systems for the best collection of web resources.

■ keyword : | Archiving Web Resources | Web Crawler | Web Snapshot Robot | Archiving System | Permanent Preservation |

I. 서 론

오늘날 우리는 디지털 정보의 홍수 속에서 살고 있다. 디지털 정보가 기하급수적으로 늘어나고 반면 급속한 폐기와 망실이 일어나고 있다. 오늘날 많은 디지털 정보를 생성하는 것에 초점을 두고 있고 보존 및 항구적인 디지털 자원들을 접근을 위해서 최근 각국에서는 민간 또는 정부 차원에서 디지털 자원들의 보존을 위한 노력들을 하고 있다.

웹은 애초에 학술정보의 공유와 유통 수단으로 만들어졌다. 그 이후 지금까지 학술 및 과학 커뮤니케이션의 수단으로서의 웹의 비중은 지속적으로 확대되고 있다. 전 세계의 연구자들은 웹을 통해서 최신 연구정보를 교환하거나 검색하고, 전자저널 같은 형태로 연구결

접수번호 : #110629-003
접수일자 : 2011년 06월 29일
심사완료일 : 2011년 08월 03일
교신저자 : 김광영, e-mail : kykim@kisti.re.kr

과물을 배포하고 획득한다[5].

웹 자원은 정보 이용자들이 가장 빠르고 손쉽게 접근할 수 있는 중요한 매체이며 과학기술 커뮤니케이션뿐만 아니라 개인 커뮤니케이션, 출판, 학술, 전자상거래 등 다양한 분야에서 활용되는 중요한 자원들이다. 하지만 이런 웹 자원들은 중요성에 관계없이 주기적 또는 비주기적으로 갱신되거나 소멸된다. 따라서 웹 자원을 수집하고 보존하는 웹 아카이빙의 중요성이 강조되고 있다. 이러한 웹 아카이빙의 관련 연구가 증가하면서, 웹 자원을 수집하기위한 웹 크롤러의 개발이 필요하게 되었고, 몇몇 웹 수집관련 프로젝트들은 수집을 위한 도구를 개발하였다[1].

웹의 역동성과 기술의존성으로 인해 웹 자원들이 계속해서 수정되고 바뀌고 삭제되고 있다. 오늘날 알고 있는 것, 즉 전자적으로 코드화 되고 기록된 것의 대부분이 영원히 사라지게 될 디지털 암흑시대로 옮겨가고 있다[9].

국외의 경우에 Internet Archive[17]는 미국 샌프란시스코에 위치한 비영리단체로서 디지털 형태로 존재하는 역사적 정보자원에 영구적으로 접근할 수 있는 “인터넷 도서관”을 구축한다는 목적을 가지고 1996년에 설립되었고 주제나 수준 등 수집대상의 범위에 제한을 두지 않고 미래를 위해 광범위하게 수집하는 정책을 유지하고 있다[6].

호주의 PANDORA, 영국의 The National Archive에서도 웹 자원에 대한 아카이빙을 수행하고 있으며, 국내의 경우에는 국립중앙도서관의 OASIS에서 웹 자원에 대한 아카이빙을 수행하고 있다. 웹을 기반으로 하는 정보자원의 보존은 웹의 고유한 매체 특성으로 인해 기존의 인쇄물을 중심으로 하는 유형 기록물의 보존에서 나타나는 것과는 현저하게 다른 문제점들이 나타내고 있다[2]. 또한 웹 자원을 작성하는 국가에 따라 다른 웹의 특징을 가지고 있기 때문에 이러한 사항을 고려한 아카이빙을 수행해야 한다. 웹 아카이빙의 절차는 선별-수집-저장-전달로 구성된다[3].

이와 같이 국내외적으로 웹 아카이빙에 관한 많은 연구들이 진행되고 있다. 따라서 본 논문에서는 웹 자원의 수집을 위해서 사용되는 기존의 웹 크롤러의 장단점을 분석하고 이것을 이용하여 가장 적합한 웹 자원 수집 도구 시스템을 연구하고 개발하는 것이다. 이에 따라 본 논문은 다음과 같이 구성되었다. 2장에서는 기존 웹 자원 수집 도구들을 분석하였고, 3장에서는 분석된 기존의 웹 자원 수집 도구의 단점을 해결할 수 있는 새로운 모델을 제시하였고, 4장에서는 제안된 모델을 이용하여 웹 아카이빙 수집 시스템을 설계 및 개발하였다. 이러한 내용은 6장에서 결론을 맺고, 향후연구에 대해 논하였다.

II. 웹 아카이빙 수집 도구 분석

웹 자원을 아카이빙하기 위한 도구를 crawler 또는 harvester라고 한다. 이는 1996년 스웨덴에서 Kulturarw3 프로젝트[11]를 진행하면서 Combine harvester에서 처음 사용되었으며, 인터넷 아카이브(Internet Archive)에서는 Alexa Crawler, 프랑스 국가도서관에서 Xyleme Crawler를 사용하였다. 이들은 크롤러의 초기적인 형태로 일반적인 색인로봇을 바탕으로 설계되었기 때문에 웹 자원 아카이빙을 위한 전용 크롤러가 필요하게 되었다.

NEDLIB(Networked European Deposit Library)에서 전용 크롤러인 NEDLIB harvester를 개발하였고[1], 이는 유럽을 중심으로 다수의 국가도서관에서 이용되었다. 그 이후에 개발된 크롤러로는 인터넷 아카이브에서 개발한 Heritrix와 웹사이트를 통째로 저장할 수 있는 HTTrack들이 있다.

국내 공공기관의 웹기록물관련 아카이빙 살펴보면 오픈소스인 Heritrix[12]를 활용하여 수집하고 WARC(Web ARChive)파일에 대한 빠른 접근을 위한 인덱스 저장기와 웹기록물 뷰어 등을 개발한 연구도 있다[4]. 또한 기존 아카이빙 도구에 관한 연구[1]에서는 NEDLIB harvester와 Heritrix를 대상으로 크롤러를 분석하였다.

[표 1]과 같이 Heritrix는 오픈소스이며 리눅스 플랫폼을 기반으로 운영이 되며 자바스크립트를 지원하며 웹 관리자 인터페이스를 제공함으로 쉽게 사용할 수가 있다[12].

HTTrack는 손쉬운 설치와 실행으로 웹사이트를 수집 할 수 있지만, 소스가 공개되어있지 않으며 일반적인 저장방식을 통한 수집 이외에 사용자들의 설정 및 수정을 제한하고 있다[13].

DeepArc는 심층 웹 자원을 수집하기 위한 관계형 데이터베이스 내용을 XML로 변환 시켜주는 것을 주목적으로 하며 자바스크립트나 플래시(flash)는 지원하지 않는다[14].

PageVault는 상용 소프트웨어로 웹 서버에서 생성되는 모든 응답(response)들을 아카이빙을 지원하며 동적/정적 웹 페이지에서 생성되는 모든 포맷(HTML, XML, PDF, zip, microsoft office formats, image, sound)을 지원한다. 그러나 자바스크립트, 플래시, 심층 웹 수집은 지원하지 않는다[15].

WGet은 프리웨어로 커맨드라인(command-line) 방식으로 HTTP, HTTPS, FTP의 콘텐츠를 다운로딩하는 도구이다[16]. 단순하게 다운로드 받고자하는 경로를 알고 있을 경우에 커맨드라인 상에서 쉽게 다운로드 할 수가 있다. 그러나 웹 사이트를 수집하는 용은 아니다.

표 1. 웹 자원 수집 도구

이름	상용	플랫폼	자바스크립트	Flash	심층 웹 수집
Heritrix	오픈소스	리눅스	○	×	불가
HTTrack	프리웨어	리눅스, 윈도우즈	×	×	불가
DeepArc	오픈소스	리눅스, 윈도우즈, 맥	×	×	가능
PageVault	상용	리눅스, 윈도우즈	○	×	불가
WGet	프리웨어	리눅스	×	×	불가

이와 같이 본 연구에서 살펴 본 웹 자원 수집 도구들은 지속적으로 변화되는 웹 기술의 대응하기 위해서는 아직 구현해야 할 기능들이 많이 남아있다. 일반 검색 시스템에서 사용하고 있는 웹 수집 크롤러와 매우 유사한 기능만을 제공하고 있다. 따라서 웹 아카이빙을 위한 수집 크롤러는 그 목적에 적합한 기능들을 지속적으로 연구 개발할 필요가 있지만 빠른 기술적 대응에 한계점이 생긴다.

III. 웹 아카이빙 수집 모델

표준 HTML 링크로 연결된 정적인 웹 페이지로 구성된 웹 자원을 수집하는 일은 비교적 쉽다. 그러나 계속 증가하고 있는 자바스크립트나 플래시 등의 스크립트나 플러그인(plug-in) 같은 기법을 사용하는 동적인 웹 페이지를 수집하는 일은 어렵다.

스크립트 실행의 결과는 웹 브라우저의 종류 등 많은 사항에 따라 달라지기 때문에 자바스크립트를 채용하는 웹 페이지는 성공적으로 수집하기는 어렵다.

플래시 역시 플러그인을 사용할 뿐만 아니라 상용포맷이기 때문에 그런 웹 페이지를 수집하기는 매우 어렵다. 또한 웹 사이트 운영자가 로봇배제표준을 사용하여 웹 사이트에 robot.txt 파일을 심어 놓은 경우에도 로봇에 의한 수집이 불가능하다.

데이터베이스를 기반으로 하는 웹 사이트들도 일반 검색엔진과 마찬가지로 수집로봇을 통해서 웹 자원을 수집하는 것은 어렵다. 이처럼 '심층 웹(deep web)'의 존재는 특히 수집로봇에 의존하는 웹 아카이빙에 심각한 문제를 제기하고 있다[6]. 즉 내부 데이터베이스를 이용한 웹 페이지 사이트와 연결이 되어 있을 경우에는 웹 자원을 수집해서 복원을 할 경우에 그 사이트가 사라지면 해당되는 웹 페이지의 내용을 알 수가 없다.

웹 공간에는 검색엔진을 통해서 자유롭게 접근할 수 있는 '표면 웹(surface web)'만 있는 것이 아니라 일반적 검색엔진에 의해 색인되지 않는 '심층 웹(deep web)'이 함께 존재하기 때문이다. '숨은 웹(hidden web)' 또는 '보이지 않는 웹(invisible web)'이라고도 일컬어지는 '심층 웹'은 대체로 표면 웹의 400-500배에 이르는 것으로 추정되고 있다[7].

심층 웹은 웹 수집로봇 자체의 기술적 한계에 의해서 생겨나기도 하고, 웹 사이트가 인증 과정이나 로봇배제 프로토콜 같은 방법으로 수집로봇의 접근을 거부하여 만들어지기도 한다. 웹 자원의 검색과 아키이빙 측면에서 가장 큰 문제가 되는 것은 전문화된 데이터베이스와 최근 급성장하고 있는 역동적 웹 사이트들이다[6].

이와 같이 현재의 웹 자원 수집을 위한 다양한 도구들이 있지만 스크립트와 플러그인 형태를 완벽하게 지

원하기가 어렵다. 또한 표면 웹보다 심층 웹 자원을 수집하기는 더 많은 시간과 어려움이 따른다. 표면 웹만을 수집할 경우에 해당 웹 사이트를 복원할 경우에 스크립트와 플래시 등이 있을 경우에 해당하는 웹 페이지의 내용을 알 수가 없다. 현재의 웹 수집 도구들은 이런 한계를 극복하기에 많은 시간과 투자가 필요하다. 따라서 이런 한계점을 극복하기 위한 대안으로 본 연구에서는 웹 페이지 사진을 찍는 스냅샷(snapshot) 로봇을 활용한 모델을 제안한다.

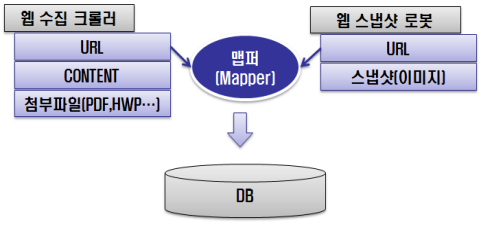


그림 1. 웹 자원 수집 모델

[그림 1]과 같이 웹 자원 수집기는 웹 문서들과 관련된 링크 파일을 수집하고 또한 해당하는 웹 페이지를 웹 스냅샷 로봇이 재방문하여 그 웹 페이지의 이미지를 캡처하여 웹 페이지 이미지를 저장한다.

저장된 이미지는 웹 자원 수집기에서 수집한 URL, 웹 문서, 링크된 파일과 스냅샷한 이미지로 관리한다. 이와 같이 구성을 할 경우에는 심층 웹을 수집하지 않아도 해당되는 웹 페이지 이미지를 활용하여 그 당시의 웹 페이지를 보여 줄 수가 있다. 또한 자바스크립트나 플래시와 같은 다양한 형태의 플러그인에도 상관없이 해당하는 웹 페이지를 이미지로 보여줄 수가 있다. 그리고 사용자가 특정 웹 문서를 찾고자 할 경우에는 수집한 문서의 내용(content)과 스냅샷 이미지 간의 맵퍼(mapper)를 이용하여 검색된 페이지의 내용과 수집된 웹 페이지 이미지를 함께 제공함으로써 그 당시의 웹 페이지 모습 등을 정확하게 제공할 수가 있다.

지금까지 기술보존, 기술 에뮬레이션(emulation), 정보마이그레이션(migration), 인캡슐레이션(encapsulation) 등과 같은 여러 가지 기법들이 제시되었으며, 이 기법들은 각각 장단점을 가지고 있는 것으로 보고되고 있다[10]. 하지만 원본 디지털자료와 신뢰성 보장이 증명된 장기 보존 기법은 아직 존재하지 않는다. 하지만 본 연구에서 제안하는 방식을 사용할 경우에 수집될 당시의 웹 사이트를 모습과 그 내용을 각각 보존하여 원래의 웹 자료를 유지하도록 하였다.

IV. 웹 자원 아카이빙 시스템 개발

본 연구에서 개발한 웹 자원 아카이빙 시스템은 웹 자원 선별, 수집, 저장 및 전달로 구성되어 있다. 특히 웹 페이지 스냅샷 사진 등과 같이 추가로 수집된 자원들도 함께 관리한다. 수집된 자원을 일반 사용자들에게 검색 서비스를 제공하기 위한 검색관리 시스템을 함께 제공한다.

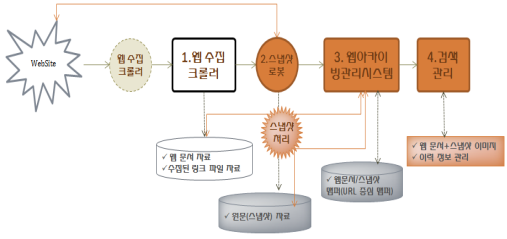


그림 2. 웹 아카이빙 시스템 구성도

[그림 2]와 같이 웹 아카이빙 시스템의 세부적인 기능들은 크게 네 개의 시스템으로 구성되어 있다.

1. 웹 수집 크롤러

웹 수집 크롤러는 일반 웹 수집 크롤러와 유사하며 선별된 수집 대상 웹 사이트의 문서뿐만 아니라 링크 파일까지 자동으로 수집하는 시스템이다. [그림 3]과 같이 수집 대상 사이트(Seed)를 관리하는 기능과 실제 HTTP로 접속하여 문서를 받아서 파싱 처리하는 HTTP/Fetch/Parser 에이전트와 파싱된 웹 문서를 일반 텍스트 문서로 변화하는 HTML2TXT 기능이 있다. 변환된 일반 텍스트 문서를 DB에 저장 및 갱신 관리하는 UpdateRobot으로 구성된다. 따라서 일반 웹 수집 크롤러는 수집하고자하는 사이트에서 웹 문서와 링크된

파일 자원들을 수집하여 위의 [그림 2]와 같이 웹 문서 자료와 수집된 링크 파일 자료를 저장하고 관리한다.

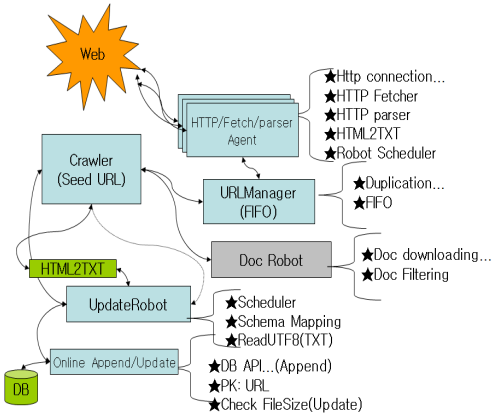


그림 3. 웹 수집 크롤러 시스템

2. 웹 스냅샷 로봇

위의 [그림 2]와 같이 웹 스냅샷(snapshot) 로봇 시스템은 웹 수집 크롤러가 성공적으로 수집한 사이트들을 대상으로 웹 수집 크롤러가 방문한 웹 페이지를 다시 방문하여 해당 페이지를 사진 촬영(snapshot)한다. 촬영한 웹 페이지의 원문(스냅샷) 자료는 DB에 저장하여 관리한다.

웹 스냅샷 로봇은 [그림 4]와 같이 각 세부 기능별로 처리를 수행한다.

1) 웹 스냅샷 로봇은 일반 웹 수집 크롤러와 유사하게 수집할 대상의 웹 페이지(Seed URL)를 중심으로 사진 촬영할 준비를 수행한다.
2) 사진 촬영할 웹 페이지들은 URL Manager에 등록하여 중복된 사이트 등을 관리하며 촬영 순서는 수집된 URL를 중심으로 FIFO 방식으로 사진 촬영할 웹 페이지를 관리한다.
3) HTTP/Fetch 에이전트는 실제 사진 촬영할 웹 페이지를 방문하여 사진 촬영을 수행한다.
4) 촬영된 이미지들은 특정 이미지 포맷으로 변환하여 저장한다.
5) 이미지 파일로 변환된 파일들은 DB에 등록되어 관리한다.

이와 같이 HTTP/Fetch 에이전트는 사진 촬영할 웹 페이지들이 모두 끝날 때까지 반복 수행한다.

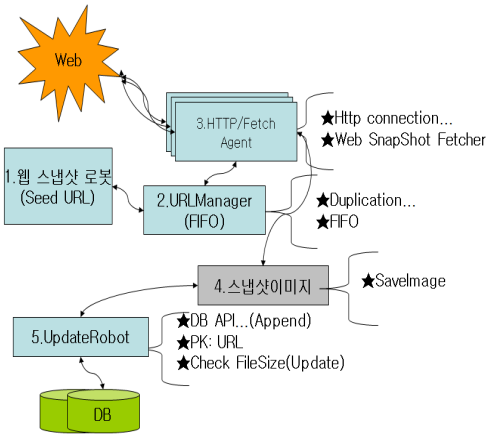


그림 4. 웹 스냅샷 로봇 시스템

3. 웹 아카이빙 관리 시스템

웹 아카이빙 관리 시스템은 일반 사용자들이 검색 및 사용자 인터페이스(UI) 서비스를 제공하기 위해서 웹 수집 크롤러에서 수집한 URL, 웹 문서 자료, 수집된 링크 파일 자료와 웹 스냅샷 로봇이 스냅샷한 원문(이미지) 자료를 URL 중심으로 맵핑시킨 맵퍼를 관리한다. 또한 웹 아카이빙 관리 시스템은 웹 수집 크롤러와 웹 스냅샷 로봇이 수집한 정보를 관리하며 웹 수집 크롤러가 수집한 웹 문서들을 일반 사용자들에게 검색서비스를 제공하기 위해서 색인(index)을 처리하며 수집된 링크 파일과 스냅샷 로봇이 제공한 이미지 파일 등을 관리 한다. 또한 관리자가 수집한 웹 페이지를 수정/삭제할 수 있는 관리 기능, 통계 기능, 웹 페이지 모니터링 등의 기능을 제공한다. 웹 아카이빙 관리 시스템은 웹에서 수집한 자료들을 영구보존 포맷인 PDF/A로 변환하여 METS 패키징으로 관리한다.

4. 검색 관리 시스템

웹 아카이빙 검색 관리 시스템은 웹 수집 크롤러와 웹 스냅샷 로봇이 수집한 자료들을 일반 사용자들이 검

색하고 열람하기 위한 시스템이다. [그림 5]와 같이 일반 사용자들이 검색을 수행하기 위해서 다음과 같은 절차로 수행된다.

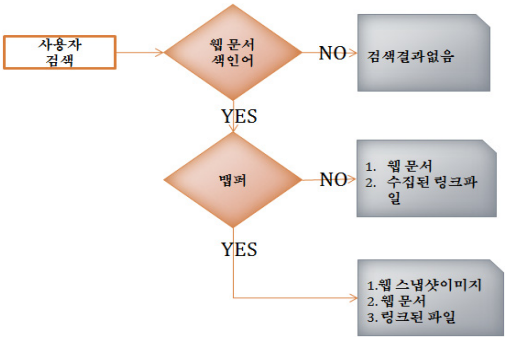


그림 5. 검색 관리 시스템 흐름도

1) 사용자가 입력한 키워드를 일반 웹 수집 크롤러가 수집한 색인된 DB에서 웹 문서들 검색한다. 만약 검색한 결과가 없으면 사용자에게 검색 결과가 없다 것을 알린다.
2) 색인어가 있을 경우에는 찾아진 웹 문서의 URL 정보를 이용하여 스냅샷 로봇이 수집한 웹 페이지 스냅샷 이미지를 맵퍼 DB에서 검색한다. 만약 검색한 이미지 결과가 없을 경우에는 웹 문서정보와 수집된 링크파일 정보만을 제공한다.
3) 맵퍼에서 찾아진 원문(스냅샷) 이미지가 있을 경우에는 웹 스냅샷 이미지, 수집된 링크 파일 정보 등을 [그림 6]과 같이 사용자 인터페이스로 구현하여 보여준다.

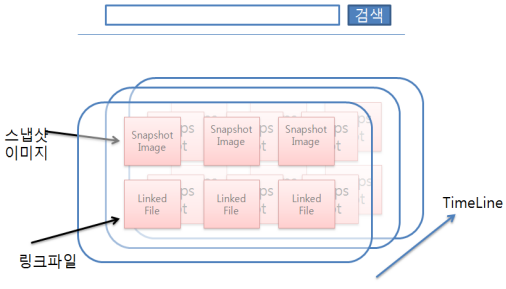


그림 6. 검색 UI 화면

[그림 6]과 같이 각 리스트들은 웹 스냅샷 이미지, 웹 문서, 수집된 링크파일 등을 함께 보여준다. 사용자가 특정 파일을 클릭할 경우에 해당되는 원본 이미지 또는 파일 정보를 보여 주게 된다.

타임라인(timeline)은 수집 주기별로 수집된 이력 정보 기능이다. 즉 특정 날짜별로 수집된 그 당시의 웹 사이트의 보존 모습을 열람할 수 있다.

본 연구에서 제공한 웹 아카이빙 시스템은 기존의 아카이빙 시스템의 문제점인 심층 웹 자원을 수집하지 못해서 해당하는 그 당시의 웹 사이트 모습을 구현할 수 없는 문제와 다양한 자바스크립트나 플래시 등의 문제점들을 웹 스냅샷 로봇을 활용하여 이런 문제점들을 해결할 수가 있었고 스냅샷한 이미지를 검색하기 위해서는 수집된 웹 문서와 스냅샷 이미지 간에 URL을 중심으로 맵퍼를 구성하여 그 연결점을 유지하도록 하였다.

V. 결론 및 제언

현재 웹 자원 수집을 위한 다양한 도구들이 있지만 스크립트와 플러그인 형태를 완벽하게 지원하기가 어렵다. 또한 표면 웹보다 심층 웹 자원을 수집하기는 더 많은 시간과 어려움이 따른다. 표면 웹만을 수집할 경우에 해당 웹 사이트가 사라지고 다시 복원할 경우에 스크립트와 플래시로 구성된 웹 페이지들의 경우에는 그 내용을 알 수가 없다. 현재의 웹 수집 도구들은 계속적으로 변화되는 웹 기술에 빠르게 대응하기 위해서는 많은 한계점들을 극복해야하며 많은 시간과 투자가 필요하다. 따라서 본 논문에서는 웹 자원의 아카이빙 수집을 위해서 사용되는 기존의 웹 크롤러의 문제점들을 분석하여 가장 효율적인 웹 자원 수집 도구를 위한 대안으로 웹 페이지 사진을 찍는 스냅샷(snapshot) 로봇을 활용한 모델을 제안하였다. 또한 사용자를 위한 검색 서비스를 위해서 웹 문서의 내용과 스냅샷한 이미지를 URL 중심으로 맵퍼로 관리하였다. 따라서 일반 사용자는 웹 문서의 내용과 함께 그 당시의 웹 페이지 이미지도 함께 열람을 할 수 있도록 하였다. 즉 스냅샷한 이미지를 검색하기 위해서는 수집된 웹 문서와 스냅샷 이미지 간에 URL을 중심으로 맵퍼를 구성하여 그 연결

점을 유지하였다. 따라서 웹 스냅샷 로봇을 활용하여 실제 웹 사이트들을 대상으로 수집을 한 결과 심층 웹, 플래시, 자바 등의 문제점들을 해결할 수가 있었다.

향후 연구과제로는 웹 스냅샷 로봇이 촬영한 이미지 기술을 확장하여 영상으로 처리할 수 있는 기술 연구와 아카이빙된 웹페이지의 장기 보존 문제에 관한 기술적인 검토와 웹 아카이빙 전략, 정책, 법적문제, 국내외 기관간의 상호협력 및 국제적 표준 참여 등의 해결도 함께 이루어져야 한다.

참 고 문 헌

[1] 이성숙, "웹 아카이빙 도구에 관한 연구", 한국정보 관리학회 학술대회, 제5권, pp.185-193, 2005.

[2] 김유승, "공공기록물 관리에 관한 법률의 제정 의의 와 개선방안", 한국기록관리학회지, 제8권, 제1호, pp.5-24, 2008.

[3] B. *Adrian, Archiving Website: a practical guide for information management professionals,* facet publishing, 2006

[4] 차승준, 정준선, 이규철, "공공기관 웹기록물 아카이빙을 위한 웹 크롤러 연구 개발", 한국정보과학회, 제25권, 제2호, pp.1-15, 2009.

[5] J. Hendler, "Science and the Semantic Web," Science 299(5606) pp.520-521, 2003.

[6] 서혜란 "웹 아카이빙의 성과와 미래 전망", 한국비블리아학술발표 제10집, pp.7-25, 2004.

[7] Bergman and K. Michael "The Deep Web: Surfacing Hidden Value," Journal of Electronic Publishing, Vol.7, No.1, 2001.

[8] A. Ball, "WEB Archiving," Digital Curation Centre, UKOLN, University of Bath, 2010.

[9] K. Terry, "The Digital Dark Ages?: Challenges in the Perservation of Electronic Information," International Preservation News No.17, pp.8-13, 1998.

[10] K. H. Lee, "The State of the Art and Practice in Digital Preservation," Journal of Research of the national Institute of Standards and Technology Vol.107, No.1, pp.93-106, 2002.

[11] P. M. Krister and A. Allan, "The Kulturarw Project - The Royal Swedish Web Archive," Electronic Library, Vol.16, No.2, pp.105-108, 1998.

[12] http://crawler.archive.org

[13] http://www.httrack.com

[14] http://bibnum.bnf.fr/downloads/deeparc

[15] http://www.projectcomputing.com/products/pageVault

[16] http://www.gnu.org/oftware/wget

[17] http://www.archive.org

저 자 소 개

김 광 영(Kwang-Young Kim) 정회원

- 2001년: 부산대학교 전자계산학과(석사)
- 2011년 : 충남대학교 문헌정보학과(박사)
- 2001년 ~ 현재 : 한국과학기술정보 연구원 선임연구원

<관심분야> : 정보검색시스템, 개인화 검색시스템, 디지털도서관, 아카이빙 시스템

이 원 구(Won-Goo Lee) 정회원

- 2000년 2월 : 한남대학교 컴퓨터공학과 (공학사)
- 2002년 2월 : 한남대학교 컴퓨터공학과 (공학석사)
- 2005년 8월 : 한남대학교 컴퓨터공학과 (공학박사)
- 2005년 2월 ~ 현재 : 한국과학기술정보연구원 정보기술연구실

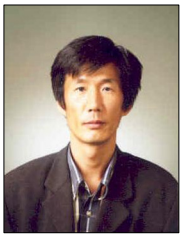
<관심분야> : 시맨틱, 디지털 아카이브, 데이터 관리

이 민 호(Min-Ho Lee) 정회원

- 2000년 : 충남대학교 대학원 컴퓨터공학과 졸업(석사)
- 2006년 : 충남대학교 대학원 컴퓨터공학과(박사수료)
- 2000년 ~ 2001년 : 데이콤 중앙연구소 연구원
- 2001년 ~ 현재 : 한국과학기술정보연구원 정보기술연구실 선임연구원

<관심분야> : 정보검색 및 추출, 정보보호, 분산시스템

윤 화 묵(Hwa-Mook Yoon) 정회원



- 2009년 2월 : 배재대학교 컴퓨터공학과(공학박사)
- 2000년 1월 ~ 현재 : 한국과학기술정보연구원 책임연구원

<관심분야> : 시맨틱기술, 콘텐츠관리, 정보검색

신 성 호(Sung-Ho Shin) 정회원

- 2000년 2월 : 경북대학교 경영학과(경영학사)
- 2002년 8월 : 경북대학교 경영학과(경영학석사/MIS전공)
- 2002년 9월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 데이터베이스 통합, 데이터품질, IS평가