제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

웹 아카이빙 시스템에 관한 연구

김광영*, 이석형*, 최호섭*, 한희준* 김진숙*
*한국과학기술정보 연구원

e-mail:{kykim, skyi, hschoe, heejun, jinsuk}@kisti.re.kr

A Study on Web Archiving System Development

KwangYoung Kim*, SeokHyoung Lee*, HoSeop Choie*, HeeJun Han*, Jinsuk KIM*
*Dept of Information Technology Research, KISTI

요 약

오늘날 디지털 정보가 기하급수적으로 늘어나고 반면 급속한 폐기와 망실이 일어나고 있다. 특히 웹 자원은 아직 수집, 보존, 활용에 대한 방안이 없어서 일정 기간의 서비스가 끝나면 사라져 버리는 문제점이 있다. 따라서 웹 자원을 수집하고 보존하기위한 웹 아카이빙 시스템이 요구되고 있다. 이러한 웹 자원들을 주기적으로 수집하여 웹 자원을 항구적인 보존과 접근을 위한 웹 아카이빙 개발이 필요하게 되었다. 따라서 본 연구에서는 웹 자원의 아카이빙 수집, 보존, 항구적인 접근을 위한 웹 아카이빙 시스템을 개발하였다.

1. 서론

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.

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[4].

국외의 경우에 Internet Archive[5]은 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].

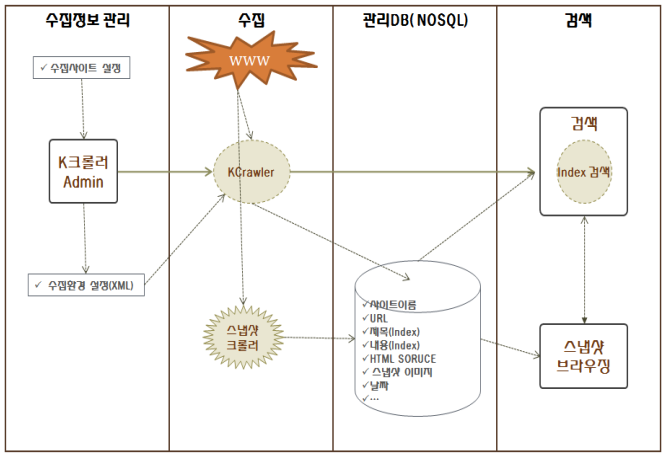
호주의 PANDORA, 영국의 The National Archive에서도 웹 자원에 대한 아카이빙을 수행하고 있으며, 국내의 경우에는 국립중앙도서관의 OASIS에서 웹 자원에 대한 아카이빙을 수행하고 있다. 웹을 기반으로 하는 정보자원의 보존은 웹의 고유한 매체 특성으로 인해 기존의 인쇄물을 중심으로 하는 유형 기록물의 보존에서 나타나는 것과는 현저하게 다른 문제점들이 나타내고 있다[2]. 또한 웹 자원을 작성하는 국가에 따라 다른 웹의 특징을 가지고 있기 때문에 이러한 사항을 고려한 아카이빙을 수행해야 한다. 웹 아카이빙의 절차는 선별-수집-저장-전달로 구성된다[3]. 국내 공공기관의 웹기록물관련 아카이빙 살펴보면 오픈소스인 Heritrix[7]를 활용하여 수집하고 WARC(Web ARChive)파일에 대한 빠른 접근을 위한 인덱스 저장기와 웹기록물 뷰어 등을 개발한 연구도 있다[8]. 이와 같이 국내외적으로 웹 아카이빙에 관한 많은 연구들이 진행되고 있다.

따라서 본 논문에서는 웹 자원의 수집하여 항구적인 접근과 보존을 위해서 웹 아카이빙 시스템을 연구하고 개발하였다. 이에 따라 본 논문은 다음과 같이 구성되었다. 2장에서는 웹 아카이빙 시스템의 워크플로우를 정의와 보존의 중요한 요소들을 정의하였고, 3장에서는 전체 아키텍처를 설계 및 개발하였다. 마지막으로 4장에서는 결론을 맺고, 향후연구에 대해 논하였다.

2. 웹 아카이빙 시스템 워크플로우

본 시스템에서 웹 자원 보존시스템의 주요한 업무처리와 기능모듈의 중심이 되는 아카이빙 워크플로우는 <그림 1>과 같이 수집정보 관리, 수집, 관리(DB), 검색 시스템으

로 4단계로 구성하였다.



<그림 1> 웹 아카이빙 시스템 흐름도

웹 수집정보 관리는 수집 크롤러의 환경 설정하는 것으로 수집 대상 사이트 설정, 수집 간격, 반복적 수집 등의 다양한 수집 크롤러의 환경을 설정할 수가 있다.

수집은 설정된 정보를 이용하여 실제 수집기가 직접 사이트에 방문하여 웹 문서, HTML 소스, 웹 이미지 등을 수집한다. 수집기는 크게 일반 웹 검색 시스템에서 사용하는 것과 유사한 크롤러와 웹 페이지의 스냅샷 사진을 촬영하는 스냅샷 크롤러로 구성된다. 일반 웹 수집 크롤러는 웹 문서, 링크 정보 등의 정보를 수집하고 웹 스냅샷 크롤러는 해당하는 웹 페이지의 이미지를 촬영하여 이미지를 수집한다.

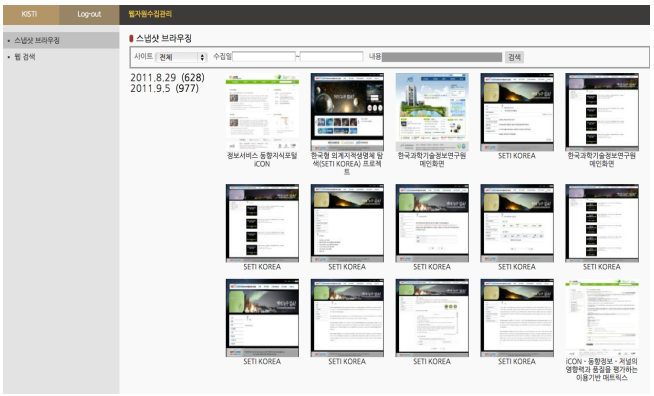
관리DB는 웹 수집기들이 수집한 내용들을 DB에 저장 및 색인 처리를 하여 관리한다. 즉 사이트 이름, URL, 웹 문서의 제목, 내용, HTML 소스코드, 스냅샷 이미지, 날짜 등의 정보 관리를 담당한다.

검색시스템은 색인된 웹 문서들을 검색과 스냅샷(shapshot)한 이미지를 브라우징을 처리할 수 있도록 한다. 브라우징 기능은 수집된 이미지를 중심으로 행당하는 웹 사이트를 탐색을 용이하게 하기 위한 것이다. 검색은 수집된 사이트의 페이지의 내용을 색인한 것을 검색하여 사용자에게 그 결과를 제공한다.

3. 웹 아카이빙 시스템

웹 아카이빙 시스템 UI는 크게 스냅샷브라우징 기능과 웹 검색 기능으로 나누어진다. 스냅샷브라우징 기능은 수집된 날짜별로 히스토리 정보를 열람할 수 있는 기능이다. 또한 특정한 날짜를 선택하여 그 날짜에 해당하는 웹 사이트를 열람할 수 있는 기능이다. 반면 웹 검색 기능은 색인된 웹 페이지의 내용들을 직접 검색을 할 수 있는 기능이다. 수집된 날짜에 상관없이 수집된 전체 웹페이지를 검색할 수가 있다.

스냅샷브라우징은 <그림 2>와 같이 수집된 날짜별로 수집한 사이트들을 브라우징할 수가 있다.



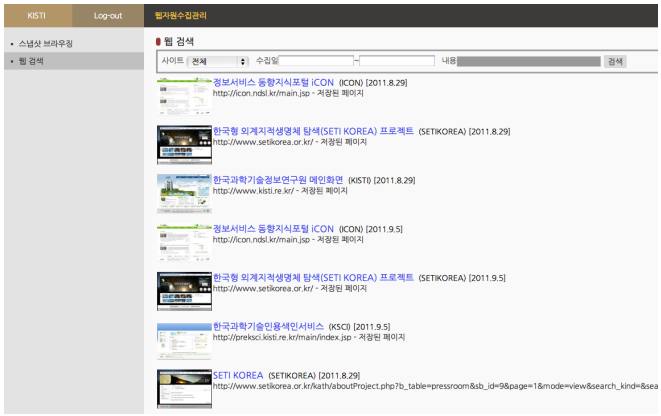
<그림 2> 웹 아카이빙 시스템 브라우징

또한 상세 내용을 보기위해서 클릭할 경우에는 <그림 3>과 같이 상세한 내용을 열람할 수가 있다. 즉 사이트 이름, URL, 제목, 내용 등을 볼 수 있으며 수집 그 당시의 웹 페이지 스냅샷한 이미지를 함께 열람할 수가 있다.

<그림 3> 상세화면

상세 화면에서는 직접 현재의 웹 사이트로 바로 가기 기능과 저장된 HTML 소스 등을 함께 제공한다. 만약에 해당하는 웹 페이지 사라질 경우에는 보존된 현재의 사이트만 열람이 가능하게 된다.

웹 검색 기능은 <그림 4>과 같이 특정 수집일과 내용으로 검색을 수행할 수가 있다. 웹 검색 기능은 수집된 웹 페이지의 내용을 직접 검색하는 기능으로 우선 수집된 웹 페이지의 내용들을 색인DB를 구성하게 된다. 구성된 색인 DB의 용어를 중심으로 검색을 수행하게 되며 수집된 날짜와는 관계없이 전체 수집된 사이트의 내용을 검색하여 해당되는 페이지만 제공한다.



<그림 4> 웹 아카이빙 검색

4. 결론

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 위한 노력으로 웹 아카이빙 시스템을 개발 연구하였다.

본 연구에서 웹 아카이빙 시스템은 수집-관리(보존)-검색(접근)과 같은 단계로 구성을 하였다. 웹 자원을 수집하기 위해서는 일반 웹 크롤러와 스냅샷 로봇을 동시에 활용하여 웹 자원을 수집하였고 수집된 자원을 보존하기 위해서는 현재 시스템은 HTML 원문, 웹 페이지 내용, 스냅샷 이미지 등을 보존하고 있다. 또한 보존한 문서들을 관리를 하기위해서 DB를 구성하였다. 사용자들의 접근을 위해서는 색인된 DB를 활용하여 브라우징과 검색을 지원하고 있다.

현재는 수집된 원문 파일을 단순하게 보존 처리하고 있지만 향후 시스템에서는 영구보존을 위한 PDF/A 포맷 등으로 마이그레이션을 처리할 것이며 보존에 대한 메타데이터를 포함하여 원본의 헤드 정보를 함께 추출하여 METS로 구성할 계획이다.

향후 연구과제로는 웹 자원의 다양한 파일들을 자동으로 보존 변화 처리 기술을 처리와 정책적인 사항들을 고려한 웹 자원 수집 대상 사이트 선정 방법 등과 같은 것들이다. 또한 수집 대상 사이트에 대한 저작권 문제 등을 고려해야할 것이다.

참고문헌

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

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

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

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

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