논문

대규모 웹 기록물의 원격수집을 위한 콘텐츠 중복 필터링 개선 연구

이 연 수*·남 성 운**·윤 대 현***

* 국가기록원 기록관리R&D센터 연구원(myth4ys@korea.com)
** 국가기록원 공업연구관
*** 한국정보화진흥원 국가정보화조정관

[국문초록]

네트워크 및 정보통신기기가 발전함에 따라 웹이 우리 일상에 미치는 영향력은 점점 더 증가하고 있다. 또한 웹 공간에서 생성되는 정보도 각 시대를 반영하는 중요한 기록물로서 그 중요성이 나날이 커지고 있다. 이에 따라 웹 정보들을 아카이빙 할 수 있는 표준화된 방법이 요구되고 있으며, 그중 한 가지가 자동화된 수집도구를 사용하여 주기적으로 수집하는 스냅샷 전략이다. 하지만 스냅샷 전략은 주기적으로 웹 콘텐츠를 수집하기 때문에 동일한 웹 콘텐츠가 중복 수집되는 문제가 있다. 또한 웹 환경에서 구현되는 복잡한 기술로 인하여 의미 없는 웹 콘텐츠가 수집될 가능성도 배제할 수 없는 실정이다.

본 논문에서는 공공기관 홈페이지 웹 콘텐츠를 스냅샷 전략으로 수집한 사례 분석을 통해서 원격 수집할 때 발생할 수 있는 콘텐츠 중복 문제들을 살펴보고, 기술 측면에 해결책을 제시하고자 한다.

주제어: 웹 기록물, 웹 콘텐츠, 원격수집, 중복제거

1. 서론

1) 연구의 목적 및 필요성

우리가 인터넷이라 부르는 월드 와이드 웹(World Wide Web)은 유통, 교육, 금융과 같은 다양한 분야에 없어서는 안 될 중요한 역할을 수행하고 있다. 이와 같은 중요성을 비추어 볼 때에 웹상의 정보자원을 보

존해야 하는 필요성도 커지고 있다. 하지만 이들을 아카이빙 할 수 있는 방안은 명확하게 제시되지 못하는 실정이다. 이처럼 웹 아카이빙에 어려움을 격고 있는 것은 웹상의 수많은 웹 콘텐츠가 복잡하게 연결되어 있기 때문이다. 또한 WAS(Web Application Sever), 운영체제, 데이터베이스 등과 같은 다양한 구조로 연결되어 있어 웹 아카이빙을 더 어렵게 만든다. 즉, 매우 다양한 환경에서 웹 콘텐츠가 존재하기 때문에 이들을 표준화하여 아카이빙하는 것을 더 어렵게 만들고 있는 것이다. 한 가지 대안 중 하나가 스냅샷(Snapshot) 방법을 이용하는 것으로 웹 서버가 가지고 있는 웹 콘텐츠를 웹 페이지의 링크 형태로 접근하여 저장하는 것이다.

이처럼 웹상의 정보자원을 보존하기 위해 수집, 저장, 서비스하는 일련의 절차를 웹 아카이빙이라 하며, 스냅샷 방법으로 웹의 일정 영역을 수집하는 것이다. 웹에 존재하는 웹 페이지나 문서와 같은 웹 콘텐츠를 스냅샷으로 수집할 때에는 이들의 주소인 URI를 발견할 수 있는 자동화된 도구를 사용한다. 이들 자동화된 수집도구를 웹 크롤러(Web crawler) 또는 웹 로봇(Web robots)이라 부른다.

그런데 수집도구를 사용하여 웹에 연결된 콘텐츠를 수집할 때에는 정책적으로 수집주기를 결정해할 필요가 있다. 왜냐하면 웹에서는 기존의 콘텐츠가 삭제되거나 또는 새로운 콘텐츠의 생성이 계속해서 발생하기 때문이다. 즉 웹의 일정 영역을 수집하여 저장하는 바로 이 순간에도 기존의 콘텐츠가 삭제될 수 있으며, 동시에 새로운 콘텐츠가 생성될 수 있다는 의미이다. 만약에 어떤 웹 사이트를 지속적으로 수집하지 않는다면, 이와 같은 변화를 반영하지 못할 것이다. 따라서 웹 콘텐츠를 아카이빙할 때는 수집주기에 따른 지속적인 수집이 중요하다고 볼 수 있다.

한편 스냅샷을 이용하여 반복적인 수집을 하는 경우에 동일한 콘텐츠가 수집될 수 있는 또 다른 문제가 발생할 수 있다. 이는 이미 저장된

대상을 다시 수집하는 것으로 불필요한 시간과 비용을 소모하게 된다. 그래서 수집도구들은 중복을 필터링하기 위한 기능을 제공하고 있다. 최근에는 WAS 서버 기반의 웹 환경에서 대규모로 웹 콘텐츠를 수집할 때에 정적인 웹 콘텐츠에서는 발생하지 않던 새로운 문제들도 나타나고 있다. 즉, 접근제한 페이지나 로그인 페이지 등과 같은 의미 없는 웹 콘텐츠들의[1] 수집이 바로 그것이다.

현재는 웹 크롤러의 필터링이 URI 중심의 필터링 방식이기 때문에 주기적으로 수집하는 웹 사이트에 의미 없는 웹 콘텐츠들이 있어서 이들이 자동 수집될 경우 이들을 필터링하는데 문제가 발생한다. 이와 같은 문제들이 공공기관 웹 기록물 수집 사례에서 발견되었으며 이를 해결하고자 WARC 저장포맷 중심의 기술 측면에 개선 방안을 다루고자 한다.

2) 연구의 방법 및 범위

공공기관 웹 사이트의 웹 기록물을 원격 수집할 때에 발생할 수 있는 문제를 확인하고 해결하기 위하여 인터넷 아카이브(Internet Archive)가 개발한 오픈소스인 Heritrix 웹 크롤러를 사용하였다. Heritrix를 선택한 이유는 Heritrix가 Java 기반의 웹 크롤러로 확장성이 뛰어나며 ISO 표준 포맷인 WARC를 기반으로 작동되기 때문이다. 또한 오픈소스 형태로 사용자가 코드를 수정할 수 있기 때문에 자체적으로 원하는 기능을 개선하거나 추가하는 것이 용이하다는 점도 작용하였다. 오픈소스의 장점은 다양한 웹 기술을 웹 크롤러에 빠르게 적용할 수 있다는 것이다. 이번 연구에 사용한 Heritrix도 수집 기능을 향상시키기 위해서 자바스크립트의 마우스 오버 이벤트[2]와 같은 함수 내부에서 작동하는 URI를

1) 접근제한 또는 로그인 페이지와 같은 해당 URI가 의도한 웹 콘텐츠가 아닌 다른 내용이 사용자에게 전달되는 경우.

찾을 수 있도록 프로그램 코드 일부를 수정하였다.

다른 종류의 웹 크롤러로는 HTTrack와 NEDLIB 등이 있는데, 이들을 사용하지 않은 이유는 구현된 언어가 플랫폼에 종속되거나 저장 포맷과 어플리케이션이 강한 결합 형태이기 때문에 표준과 확장성 측면에서 문제가 있기 때문이다.

웹 콘텐츠 저장포맷은 2009년에 웹 기록물 표준[3]으로 채택된 WARC 포맷을 사용하였다. WARC 포맷은 수집과정에서 발생하는 HTTP 프로토콜 정보를 웹 콘텐츠와 같이 저장하여 관리할 수 있게 한다. 다시 말하면, 수집환경 정보를 웹 콘텐츠와 함께 저장하여 관리토록 함으로써, 수집 당시의 웹 콘텐츠들이 위치했던 웹 서버에서의 URI 정보와 연계하여 원래 모습대로 재현하여 서비스를 제공하는 것을 가능하게 해 준다는 것이다.

원격수집 대상은 각각의 공공기관을 대표하는 메인 URI로 하였으며, 수집하는 웹 콘텐츠들은 메인 URI 영역에 포함된 모든 웹 자원을 대상으로 하였다. 다만, 원격수집도구의 기술적 한계로 발견하지 못하는 웹 콘텐츠들은 부득이하게 수집대상에서 제외 되었다. 또한 원격수집도구를 사용하는 스냅샷 전략은 웹 서버의 네트워크 트래픽에 영향을 줄 수 있으므로 각 기관의 로봇 정책[4]을 확인하여 준수하는 방향으로 실행하였다.

마지막으로 위와 같은 방식의 원격수집을 할 때에 발생할 수 있는 제반 문제를 확인하기 위하여 저장된 WARC 포맷의 로그파일을 분석하였으며, 문제가 발생하는 웹 콘텐츠들에 대해서는 해당 홈 페이지에 직접 접근하여 원인을 확인하고 규명하는 방식을 취하였다.

---

2) 웹 페이지에서 특정 영역에 마우스 커서가 위치했을 경우 발생하는 액션.
3) Web ARChive file format의 약자로 웹 콘텐츠를 저장하기 위해서 ARC 포맷을 확장한 파일 포맷.
4) 자동화된 웹 크롤러에 의한 무분별한 수집을 막기 위해서 수집 제한 웹 콘텐츠를 웹 사이트 도메인의 robots.txt 안에 명시한 것.

3) 선행연구

웹 아카이빙은 HTTP 프로토콜 기반 위에서 수행된다. HTTP 프로토콜은 클라이언트와 서버가 네트워크 통신에 있어 지켜야 할 규약을 정의하고 있다. 서버와 클라이언트는 HTTP Header를 사용하여 통신을 하게 되며, HTTP Header에 웹 자원의 변경여부를 클라이언트가 인지할 수 있게 Last-modified 필드와 E-Tag 필드가 정의되어 있다. 두 필드는 웹 자원의 변경여부를 클라이언트인 웹 크롤러가 판단하여 중복을 처리할 수 있게 해 주는 기본 데이터가 된다. 프로토콜에서 정의한 데이터를 사용하여 중복을 발견하는 방식은 표준으로 정의된 것을 사용하는 것으로 다음에 언급할 Heritrix에도 이와 같은 기능이 구현되어 있다.

저장된 데이터의 중복을 제거는 방식은 다양하며, 쉽게는 물리적으로 저장된 데이터의 비트를 비교하여 발견할 수 있다. 하지만 웹 콘텐츠와 다양한 메타데이터가 하나의 논리적 단위로 저장되는 웹 기록물 아카이빙 포맷인 WARC에는 적용하기 힘들며, 처리하는데 많은 시간과 비용이 든다. 그렇기 때문에 Heritrix에서는 Deduplicator 모듈을 개발하여 주기적인 수집 환경에서 중복을 제거할 수 있는 기능을 제공하고 있다. 중복 제거는 두 가지 방식으로 이루어지는데 앞서 설명한 HTTP Header를 사용하는 방식과 웹 자원을 다운로드 한 후 필터링 하는 방식이다. 첫 번째 방식은 HTTP Header의 유형, 용량, 수정 날짜 등 다양한 메타데이터를 사용하여 웹 콘텐츠를 수집하기 전에 중복을 판단하는 것이다. 이와 같이 웹 콘텐츠를 수집하기 전에 중복을 판단하는 것은 서버의 로드를 덜어주고 불필요한 저장을 사전에 막을 수 있다는 장점이 있다. 하지만 판단 기준이 되는 HTTP Header는 서버의 구현에 따라 누락 혹은 잘못된 정보가 포함될 수 있다. 이러한 문제 때문에 두 번째 방식인 웹 콘텐츠를 다운로드 한 후 판단하는 방식과 병행하여 사용하고 있다. 두 번째 방식은 해시함수를 사용하여 수집 대상 URI와 스트리

밍 데이터를 일련의 키 값으로 만들어 저장하는 것인데, 동일한 URI가 수집될 때 이전에 저장된 URI의 해시 값을 비교하는 것이다. 두 번째 방식은 중복되는 웹 콘텐츠를 대부분 제거할 수 있다. Heritrix에서는 위의 기능들을 Lucene과 Berkely DB 라이브러리를 사용하여 중복을 제거할 수 있게 구현되어 있다.

웹 크롤링에서 중복을 제거하는 방식은 대부분이 위에서 설명한 구조를 따르고 있는데, 다만 구현 방식의 차이만 있을 뿐이다. 하지만 애플리케이션 단계에서 구현된 필터링을 적용하지 않고 웹 콘텐츠를 수집할 경우 이를 처리하기 위해서는 추가적인 필터링이 요구된다. 본 논문에서는 애플리케이션 단계에서 처리되지 않고 저장된 웹 아카이빙 WARC 포맷에 존재하는 웹 콘텐츠 중복을 제거하고자 한다. 공공기관 웹 기록물 수집 사례에서 발생했던 웹 콘텐츠 중복을 대상으로 하였으며, 중복되어 저장된 로그인 페이지, 접근 제한 페이지와 같은 의미 없는 웹 콘텐츠를 효율적으로 제거하는 방안도 같이 제시하였다.

2. 이론적 배경

1) 웹 기록물의 수집

여기서 다루는 웹 기록물은 웹 사이트 표면에 링크 형식으로 연결된 표면 웹 콘텐츠들을 말한다. 그리고 표면에 연결되지 않은 웹 콘텐츠들도 있는데 이들은 심층 웹이라고 한다. 표면에 연결된 웹 콘텐츠의 수집을 위해 원격에서 자동으로 수집처리를 할 수 있는 로봇인 웹 크롤러를 사용하며, 이 로봇은 표면으로부터 연결된 링크들을 따라가며 웹 자원을 수집한다. 하지만 표면에 연결되지 않는 심층 웹은 데이터베이스

형태로 존재하는데 웹 로봇이 접근할 수 없어 원격수집이 불가능하다. 이들 2가지 유형의 웹 기록물 특징은 다음과 같다.

- 표면 웹 : 표면 웹은 웹 브라우저를 통해 접근할 수 있는 웹 사이트 표면에 링크 형식으로 연결된 자원을 말한다. 이렇게 연결된 표면 웹의 유형으로는 텍스트, 사무용 전자파일, 영상, 이미지 등과 같은 다양한 포맷들이 존재한다. 이들은 웹 로봇과 같은 원격수집도구를 사용하여 연속적으로 연결된 웹 콘텐츠 형태로 자동으로 수집하는 것이다.

- 심층 웹 : 심층 웹은 웹 사이트 내부에 숨겨진 자원으로써 대부분 데이터베이스 형태로 존재한다. 즉, 데이터베이스에 저장되어 있기 때문에 이들 심층 웹 자원에 접근하기 위해서는 쿼리(Query Language)를 통해서만 가능하며, 표면 웹을 원격수집하는 대부분의 웹 로봇으로는 접근이 불가능하다. 따라서 데이터베이스에 대한 환경설정 정보와 데이터를 XML 형태로 수집하는 방법이 연구되고 있다. 이들 심층 웹을 수집하기 위한 연구는 세계적으로 진행되고 있는데, 대표적으로 브라이트플래닛(BrightPlanet)과 심층 웹 연구기관(Deep Web Research)을 들을 수 있다.

2) 웹 기록물의 저장포맷

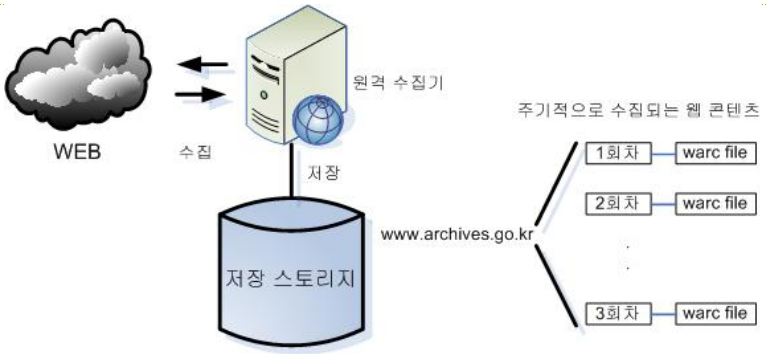
웹 기록물의 저장포맷은 XML, 데이터베이스 등 여러 유형의 포맷이 고려될 수 있다. IIPC(International Internet Preservation Consortium)는 기존의 ARC(ARChive) 포맷을 확장한 WARC를 국제표준 ISO에 제안하였으며 2009년에 표준으로 채택되었다. WARC 포맷은 다음과 같은 8가지 유형으로 구조화되어 있으며, 원격수집되는 웹 콘텐츠들은 바이너리 데이

터를 저장할 수 있는 Response 혹은 Continuation 유형 등에 저장하게 된다.

① Warcinfo : WARC 포맷에 대한 정보를 표현하는 형식
② Response : 웹 서버로부터 수집한 웹 콘텐츠 데이터를 저장하는 형식
③ Resource : 오프라인의 자원을 저장하기 위한 메타데이터 형식
④ Request : HTTP 프로토콜의 request 정보를 저장하기 위한 메타데이터 형식
⑤ Metadata : 저장된 표면 웹 콘텐츠에 대한 추가 정보를 표현하기 위한 형식
⑥ Revisit : 수집 대상이 이미 저장되어 있을 경우 사용되는 메타데이터
⑦ Conversion : 저장된 표면 웹 콘텐츠를 다른 포맷으로 저장할 때 사용되는 메타데이터
⑧ Continuation : 표면 웹 콘텐츠를 논리적으로 나누기 위한 메타데이터

3) 웹 콘텐츠의 중복 원인

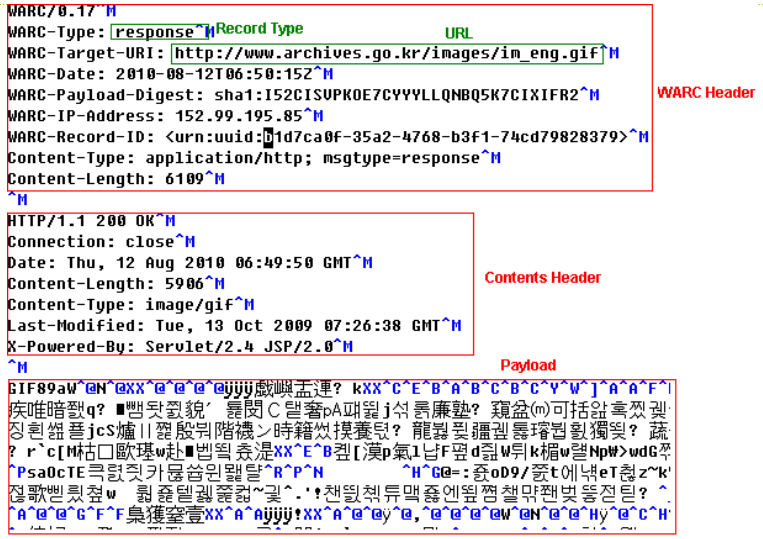
웹 기록물의 수집은 웹 사이트의 도메인 영역 전체를 대상으로 하며, 실제로 원격수집을 수행하는 웹 크롤러는 웹 콘텐츠의 URI를 기준으로 수집한다. <그림 1>은 수집대상 도메인이 www.archives.go.kr인 사이트의 표면 웹 기록물을 주기적으로 수집하였을 때 스토리지에 WARC 포맷들이 수집 회차별로 저장되는 모습이다. 이처럼 한 사이트에 대해서 주기적으로 수집할 경우에 변경되지 않았거나, 또는 의미 없는 동일한 콘텐츠들이 수집된다.



<그림 1> 스토리지에 수집 회차별로 저장된 WARC 파일들

주기적으로 수집되는 웹 콘텐츠들을 담고 있는 WARC 파일의 내부 구조는 <그림 2>와 같다. 그림을 보면 WARC는 WARC Header, Contents Header, Payload의 세 부분으로 이루어져 있다. WARC Header는 웹 콘텐츠를 수집할 때 이들을 관리하기 위해서 수집환경 정보와 UUID[5] 그리고 해시 값을 저장한다. 웹 콘텐츠 수집은 HTTP 프로토콜을 통해 이루어지는데, Contents Header는 바로 이 HTTP 헤더정보를 저장한다. 실제 웹 콘텐츠의 바이너리 데이터는 <그림 2>에서 보는 바와 같이 Payload에 저장된다. 영역 첫 줄의 GIF89a는 이미지의 헤더정보이며 뒤쪽의 데이터는 실제 저장된 웹 콘텐츠 바이트 값이다.

5) Universally Unique IDentifier로 분산된 네트워크 환경에서 디지털 객체를 구분하기 위해 사용되는 값.



<그림 2> WARC 포맷의 내부 구조

3. 웹 기록물의 원격수집 사례 및 중복 유형 분석

1) 원격수집된 웹 콘텐츠의 중복 유형

공공기관 웹 사이트에서 원격수집한 웹 콘텐츠는 두 가지 유형으로 구분할 수 있는데 동적 웹 콘텐츠와 정적 웹 콘텐츠이다. 대부분 공공기관에서 운영하는 웹 사이트는 양방향 통신을 할 수 있는 구조로 구축되어 있다. 이 조건을 만족하기 위해서는 기술적으로 WAS 서버와 데이터베이스를 사용하게 된다. 이러한 환경 특성 때문에 수집한 웹 콘텐츠 유형은 대부분 동적 웹 콘텐츠 유형이 차지하게 된다.

이들 2가지 유형의 웹 콘텐츠들에 대한 중복 사례를 보면, 먼저 정적

콘텐츠의 경우에 <그림 3>에서와 같이 URI는 다르지만 동일한 이미지를 갖는 웹 콘텐츠가 반복해서 수집되었던 것을 확인할 수 있다. 그림에서 보듯이 이미지는 하나이지만, 이 이미지를 보여주는 URI는 다음과 같이 서로 다른 2개가 존재하는 것이다.

- http://www.archives.go.kr/archivesdata/upFile/mboard/20091209073749718.jpg
- https://www.archives.go.kr/archivesdata/upFile/mboard/20091209073749718.jpg

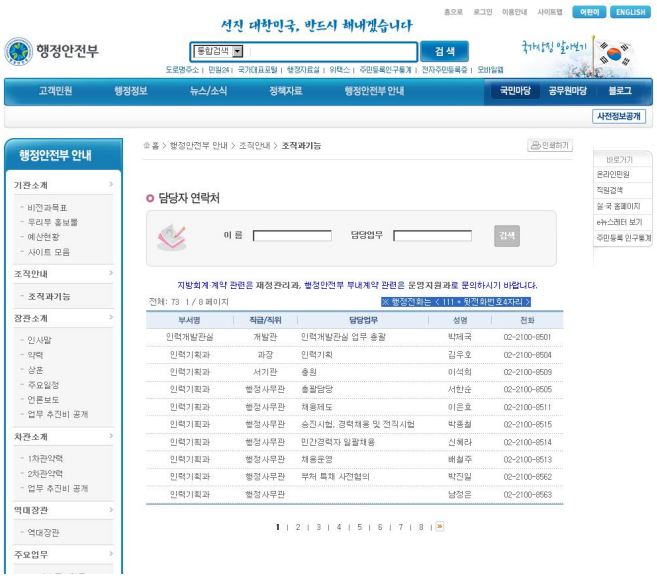
이 URI를 보면 프로토콜 부분이 HTTP와 HTTPS로 표현되어 서로 다르게 인식된다는 것이다. 즉, 동일한 이미지인데 URI가 다르게 표현되어 반복적으로 수집된 사례인 것이다.

<그림 3> 하나의 동일한 이미지가 2개의 URI로 반복 수집된 사례

다음은 동적 웹 콘텐츠의 경우에 그림 4에서와 같이 웹 프로그램으로 생성되는 동일한 웹 콘텐츠가 여러 개의 서로 다른 URI로 반복 수집되는 것을 확인할 수 있다. 그림에서 (a)의 경우는 동일한 웹 콘텐츠가 여러 개의 URI에서 생성되지만, 관련 내용이 함께 수집되어 의미가 있

다고 할 수 있다. (b)의 경우는 웹 콘텐츠의 URI를 발견하였지만 접근이 제한되어 있어서 실제 수집해야 할 웹 콘텐츠 대신에 접근을 허락하기 위해 연결된 실명인증 페이지가 반복적으로 수집된 것이다. 즉, 해당 웹 콘텐츠들의 실제 내용에 접근하기 위해서는 실명인증을 해야만 하기 때문에 수집 대상 웹 콘텐츠 대신 실명인증 페이지가 수집된 것이다. 이 둘의 차이는 다음과 같다.

(1) a 사례: 동일한 웹 콘텐츠가 다양한 URI에 반복적으로 수집된 경우
(2) b 사례: 수집 대상 웹 콘텐츠 대신 수집 대상이 아닌 로그인 또는 접근제한 페이지와 같은 의미 없는 웹 콘텐츠를 수집한 경우



(a) 사례

(b) 사례

<그림 4> 웹 프로그램에 의해 자동 생성되는 동일한 웹 콘텐츠 사례

[표 1]은 실제로 수집된 <그림 4>의 (a)와 (b)에 대한 여러 개의 상이한 URI와 해당 웹 콘텐츠의 해시 값을 보여준다. (a)의 경우는 실제로 확인된 상이한 URI가 14개인데 그중에 대표적인 3개이며 해당 URI의 웹 콘텐츠의 해시 값은 동일한 것으로 나타났다. (b)의 경우도 여러 개의 상이한 URI가 존재하였으며 이들 상이한 URI의 웹 콘텐츠 해시 값 또한 마찬가지로 동일하였다.

[표 1] <그림 4>의 동적 웹 콘텐츠에 대한 상이한 여러 개의 URI와 동일한 해시 값

	상이한 여러 개의 URI	동일한 해시 값
그림 4 (a)	(1)https://www.mopas.go.kr/gpms/ns/mogaha/user/userlayout/adminaddress/kor/adminAddressList.action?upDeptCode=&adminAddressBean.selectedTeam=1311056&tempSelectedTeam=54433&searchName=&searchDesc=&currentPage=1 (2)https://www.mopas.go.kr/gpms/ns/mogaha/user/userlayout/adminaddress/kor/adminAddressList.action?currentPage=1&upDeptCode=&adminAddressBean.selectedTeam=1311056&tempSelectedTeam=54433&searchName=&searchDesc= (3)http://www.mopas.go.kr/gpms/ns/mogaha/user/userlayout/adminaddress/kor/adminAddressList.action?currentPage=1&upDeptCode=&adminAddressBean.selectedTeam=1311056&tempSelectedTeam=54433&searchName=&searchDesc=	SHA1:ZUP7FMO5ZBO72CVOHRQV7KOS2OXVBQ42
그림 4 (b)	(1)http://www.pps.go.kr/gpms.tdf?a=user.realname.RealNameCheckApp&c=1001&mc=P_09&reUrl=&mc=P_09&reUrl=/bar03a.gif (2)http://www.pps.go.kr/gpms.tdf?a=user.realname.RealNameCheckApp&c=1001&mc=P_09&reUrl=&mc=P_09&reUrl=/9999/99/99 (3)http://www.pps.go.kr/gpms.tdf?a=user.realname.RealNameCheckApp&c=1001&mc=P_09&reUrl=&mc=P_09&reUrl=/Msxml2.XMLHTTP	SHA1:AC2N2Q7WTNZI4CF7FKOZRNXWP7YVRHFX

두 사례 중 (b)의 사례는 실제 수집대상인 웹 콘텐츠를 A라고 할 때에 접근권한이 없는 이들에게는 A대신 B라는 의미 없는 페이지가 연결되기 때문이다. 따라서 접근권한이 없는 웹 크롤러가 의미 없는 페이지들을 반복적으로 수집하게 되는 것이다. 즉, 똑같은 중복이지만 수집할 때 필터링해야 하는 사례라고 볼 수 있다.

[표 2]는 공공기관별 웹 기록물 수집 성공률로 중소기업청, 병무청,

국방부, 방위산업청의 중복이 가장 많이 발생하였으며 나머지 기관들은 20% 이하의 중복이 발생한 것을 알 수 있다. [표 2]가 의미하는 것은 수집 대상 웹 사이트의 중복 발생 비율이 높을수록 웹 콘텐츠 접근을 위한 세션 값이 더 많이 요구된다는 것이다. 즉, 앞서 살펴본 동일한 웹 콘텐츠가 임의의 URI에서 반복적으로 수집될 수 있다. 그렇기 때문에 이를 해결하기 위해서는 개선된 필터링 방안이 요구되며, 기존의 수집 전, 후 필터링 방식을 개선할 필요가 있다.

[표 2] 공공기관 웹 기록물 수집 결과[6)]

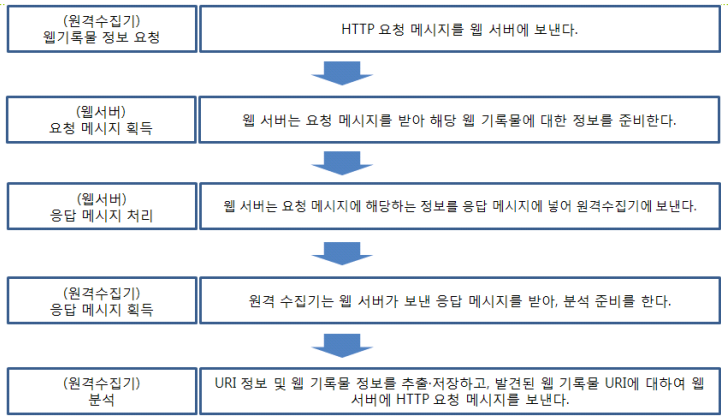
기관	전체용량(Byte)	발견된 기록물(개)	중복 제외 기록물(개)	수집성공 기록물 용량 (Byte)	성공률(%)
감사원	445,542,821	3,646	3,607	445,429,939	98.93%
검찰청	6,398,344,031	1,820,978	1,738,658	6,397,846,408	95.48%
경찰청	2,430,262,053	41,459	41,015	2,429,640,180	98.93%
공정거래위원회	4,789,910,482	428,496	400,867	4,735,304,214	93.55%
관세청	66,654,083,148	722,845	719,809	66,652,122,126	99.58%
교육과학기술부	239,799,503	2,770	2,744	239,627,664	99.06%
국가기록원	15,905,533,804	258,364	226,995	15,614,352,214	87.86%
국가보훈처	19,490,315,700	188,771	182,762	19,478,951,604	96.82%
국가인권위원회	11,687,132,067	157,453	156,537	11,686,031,850	99.42%
국무총리실	94,619,581	2,889	2,805	94,430,511	97.09%
국민권익위원회	371,064,714	1,240	1,040	211,441,778	83.87%
국방부	1,705,153,418	30,424	23,638	1,703,649,922	77.70%
국세청	4,385,052,023	32,047	28,448	4,382,546,074	88.77%
금융위원회	31,259,656,728	103,595	103,542	31,259,457,790	99.95%
기상청	134,265,622,115	1,042,444	1,026,333	134,243,012,046	98.45%
기획재정부	24,913,888,675	320,186	319,602	24,901,380,147	99.82%
노동부	1,647,059,971	10,830	10,420	1,645,119,339	96.21%
문화재청	28,377,399,109	350,078	337,952	28,371,985,088	96.54%

6) 국가기록원에서 연구 개발한 웹 아카이빙 테스트 베드를 사용하여 공공 기관 웹 사이트를 2009년 12월부터 2010년 11월까지 수집한 결과.

문화체육관광부	28,038,692,655	136,650	125,228	28,031,608,200	91.64%
방송통신위원회	18,594,719,440	268,074	223,803	18,479,530,425	83.49%
방위산업청	1,998,879,665	7,266	5,756	1,996,748,232	79.22%
법무부	30,494,817,672	811,846	810,838	30,492,585,155	99.88%
법제처	4,063,130,216	69,246	59,302	4,059,346,542	85.64%
병무청	1,689,724,448	27,172	19,458	1,689,530,737	71.61%
보건복지부	54,207,784,049	243,619	240,850	54,200,910,401	98.86%
산림청	213,466,919,560	2,854,640	2,843,662	213,309,018,608	99.62%
소방방재청	1,620,712,527	15,719	13,120	1,619,850,420	83.47%
여성부	1,877,720,746	2,107	2,069	1,877,601,479	98.20%
외교통상부	827,526,616	6,919	6,811	827,436,757	98.44%
조달청	1,152,473,363	18,916	16,972	1,150,375,681	89.72%
중소기업청	13,406,525,884	872,209	612,094	13,405,933,939	70.18%
지식경제부	17,723,600,888	236582	223,041	17,697,405,952	94.28%
통계청	29,092,141,328	282920	260,807	29,082,491,663	92.18%
통일부	21,039,078,399	427,057	426,610	21,038,705,001	99.90%
특허청	3,564,429,922	17,026	16,320	3,561,869,427	95.85%
해양경찰청	27,333,353,766	158,893	155,277	27,306,929,656	97.72%
행정안전부	59,179,522,312	535994	498,182	59,156,311,969	92.95%
행정중심복합도시건설청	29,393,367,130	716638	663,076	29,384,133,906	92.53%
환경부	8,605,788,466	170,806	170,427	8,605,385,482	99.78%

2) 웹 크롤러의 중복 필터링

표면 웹 기록물을 원격수집하는 웹 크롤러는 URI를 기준으로 수집하며, <그림 5>와 같이 5단계의 세부 작업 프로세스를 거쳐 처리한다. 그림에서 중복을 판단하는 단계는 응답 및 메시지 획득 단계로 웹 서버로부터 받은 프로토콜 정보 및 콘텐츠를 가지고 중복여부를 가린다. 중복을 필터링하는 방식은 웹 크롤러에 따라 다양한 모듈을 사용하여 구현될 수 있는데, 여기서는 원격수집시 사용한 Heritrix 웹 크롤러를 대상으로 중복 필터링을 분석하였다.



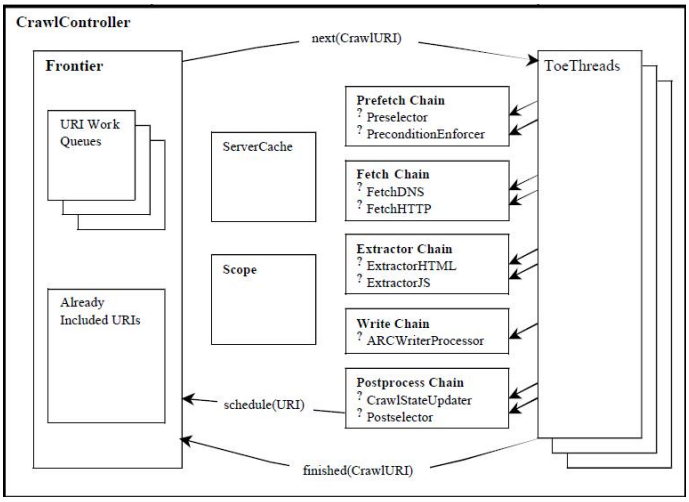
<그림 5> 원격수집 웹 크롤러의 세부 작업 프로세스

Heritrix 웹 크롤러는 도메인 영역에 포함된 모든 웹 자원을 수집하는데 적합하도록 구현되어 있다. <그림 6>의 내부구조를 보면 수집대상을 관리하는 Frontier와 입출력(I/O)작업을 하는 ToeThreads 컴포넌트들로 되어 있으며, ToeThreads 컴포넌트는 전처리, 추출, 저장, 후처리 등의 프로세서로 세분화되어 있다. 여기서 원격수집하는 웹 크롤러에서 웹 콘텐츠의 중복 필터링은 HTTP Header값을 이용하는 전처리 방식과 바이너리 데이터 해시 값을 이용하는 후처리 방식으로 구분한다. 전처리 방식은 Fetch Chain 프로세서에서 이전에 수집된 웹 콘텐츠 HTTP 프로토콜의 E-Tag 값과 Last-modified 값을 이용한다. 후처리 방식은 Postprocess Chain 프로세서에서 이전에 수집한 웹 콘텐츠의 바이너리 데이터 해시 값을 이용한다. 이 중에서 처리가 빠른 것은 전처리 방식으로 웹 콘텐츠들을 웹 서버에서 다운로드하지 않고 중복을 판별하기 때문이다. 하지만 웹 서버에 따라 E-Tag 값을 생성하지 않거나 생성하더라도 값이 다를 수 있어 정확성이 떨어진다. 반면에 후처리 방식은 웹 콘텐츠의 바이너리 데이터의 해시 값을 계산하여 직접 비교하는 것으로 정확하지만 속도가 떨어진다.

지금까지 살펴본 Heritrix 웹크롤러의 중복 필터링 방식은 수집 전과 후로 필터링을 할 수 있으며, 웹 콘텐츠를 웹 기록물 형태로(WARC 포맷) 저장하기 전에 필터링하는 구조이다. 앞서 논의한 문제를 이와 같은 구조가 수용하기 위해서는 URI 중심의 필터링 방식에 웹 콘텐츠 내용 위주의 필터링이 필요한데 조건은 다음과 같다.

1) 다양한 URI에서 무작위로 발생하는 의미 없는 웹 콘텐츠를 필터링할 수 있어야 함
2) 웹 아카이빙 WARC 포맷 내에 중복된 웹 콘텐츠[7]의 중복을 제거할 수 있어야 함

※ 수집이 완료된 웹 콘텐츠 중 중복 필터링이 미적용되어 WARC 포맷으로 저장된 것들.



<그림 6> Heritrix 웹크롤러의 내부 구조[8]

7) 웹 기록물에 저장된 웹 콘텐츠의 비트 구조가 동일한 객체들.
8) 출처 : the 4th International Web Archiving Workshop, 「Introduction to Heritrix」.

4. 웹크롤러의 중복 필터링 기능 개선

기존의 중복 필터링 방식은 URI가 동일할 때 HTTP 프로토콜 정보 또는 바이너리 해시 값을 이용하여 중복여부를 판단한다는 것을 앞에서 확인하였다. 그러나 이 방식은 URI 중심의 필터링으로 인해 무작위로 발생하는 의미 없는 웹 콘텐츠를 필터링하는데 한계가 있다. 본 장에서는 이를 해결하기 위해 웹 콘텐츠 중심의 필터링을 제시하고자 한다.

1) 원격 수집시 사전 필터링

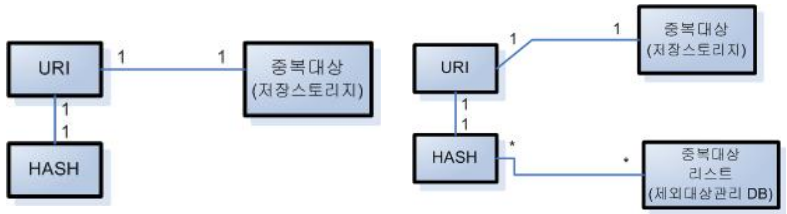
원격 수집시 사전 필터링은 최근 수집된 웹 기록물인 n+1차에[9] 대해서 중복된 웹 콘텐츠를 필터링하는 것이다. Heritrix 웹 크롤러의 기존 필터링 방식은 <그림 7> (a)에서와 같이 수집하는 URI와 이미 스토리지에 저장되어 있는 동일한 URI의 비교를 통해 중복여부를 판단하는 것이다. 그림 (a)을 보면 수집하는 웹 콘텐츠의 URI를 이미 저장되어 있는 중복대상 웹 콘텐츠 URI와 1:1 매핑하는 것으로, 중복여부의 비교 대상이 동일한 URI인 것이다. 이와 같은 방식은 동일한 URI를 가지고 웹 콘텐츠의 중복여부를 판단하기 때문에 빠르게 필터링할 수 있다. 하지만 앞에서 살펴본 것처럼 의미 없는 웹 콘텐츠의 경우에는 URI가 서로 다르기 때문에 이와 같은 방식을 적용하는데 한계가 있다.

이를 개선하기 위해 웹 콘텐츠의 해시 값을 비교하여 중복여부를 판단하는 방안을 제시하고자 한다. 이 경우에도 이미 저장되어 있는 모든 웹 콘텐츠를 비교 대상으로 한다면 강력한 컴퓨팅 파워를 필요로 할 것

September 2004.

9) 이전에 수집된 웹 기록물을 n차라고할 때 최근 수집된 기록물을 n+1차라고 정의한다.

이다. 따라서 <그림 7> (b)에서와 같이 수집에서 제외할 의미 없는 웹 콘텐츠들을 웹 아키비스트가 사전에 분석하여 확인하고, 이들의 해시 값을 새로운 중복대상 리스트로 작성하여 관리하는 것이다. 그리고 원격수집을 할 때에 콘텐츠의 해시 값과 이들 리스트를 비교하여 동일여부를 판단한다면 의미 없는 웹 콘텐츠의 반복적인 수집을 효과적으로 필터링할 수 있을 것이다. 단, 수집에 제외할 대상이 증가할 경우 수집 속도에 영향을 줄 수 있다. 그 이유는 수집 프로세서가 수집 대상인 웹 콘텐츠의 해쉬 값과 <그림 7> (b)의 중복대상 리스트에 있는 웹 콘텐츠 해시 값을 비교하는 계산양이 증가하기 때문이다. 따라서 수집 제외 대상을 신중하게 선정하여 리스트를 관리하여야 할 것이다.



(a) 기존방식 (b) 개선된 방식

<그림 7> Heritrix 웹그롤러의 필터링 방식

2) 원격수집 후 내부 필터링

WARC 포맷 파일 내에 존재하는 의미 없는 웹 콘텐츠를 제거하는 것이 내부 필터링 방식이다. 수집되는 웹 콘텐츠의 실제 데이터는 WARC의 Response 유형의 한 형태로 저장된다. 이때 WARC 포맷을 손상시키지 않고 Response 유형으로 저장된 의미 없는 웹 콘텐츠들을 제거하기 위해서는 Revisit 유형을 사용하여야 한다. 그런데 Response 유형을 Revisit 유형으로 변경시키는 것은 메타데이터 변경을 필요로 한다. [표 3]은

Response 유형의 메타데이터이다. 표에서 Response 유형의 메타데이터는 10가지의 유형으로 구분하며, 이들 중에서 'WARC-' 형태로 시작하는 메타데이터 유형들이 웹 콘텐츠들에 대한 메타데이터이고, 나머지 Content-Type과 Content-Length는 웹 서버로부터 수집한 웹 콘텐츠의 MIME 유형과 바이트 크기를 저장하는 부분이다.

[표 3] WARC 포맷의 Response 유형의 필드들

구분	설명
WARC/0.17	- WARC 포맷 버전
WARC-Type	- WARC 헤더 유형
WARC-Target-URI	- 수집되는 웹 콘텐츠들의 URI
WARC-Warcinfo-ID	- Warcinfo 헤더의 UUID 값
WARC-Date	- 저장된 날짜
WARC-Payload-Digest	- 바이너리에 대한 메시지 다이제스트 값
WARC-IP-Address	- 수집대상 웹 기록물의 IP
WARC-Record-ID	- 수집되는 웹 콘텐츠의 URI에 대한 UUID
Content-Type	- 수집되는 웹 콘텐츠의 MIME 유형
Content-Length	- Content Header와 Payload의 전체 바이트 크기

[표 4]는 Revisit 유형의 메타데이터이다. Response 유형에 없는 WARC-Refers-To와 WARC-Profile 메타데이터가 추가되어 있다. WARC-Refers-To는 WARC 포맷에 저장된 웹 기록물을 참조한다. WARC-Profile은 WARC 필드와 WARC Record 블록에 대한 처리방식을 결정하기 위해 사용된다. Revisit 유형의 본래 목적은 재수집되는 웹 콘텐츠의 URI가 동일할 경우에 Response 유형으로 수집하지 않고 Revisit 유형으로 수집하여, WARC 포맷 내에서 동일한 웹 콘텐츠들을 논리적으로 연결하는데 사용한다.

필터링 절차를 수행하지 않고 저장된 웹 아카이빙 WARC 파일과 WARC 내에 의미 없는 웹 콘텐츠를 제거하기 위해서는 WARC의 Revisit 유형을 사용해야 한다. 하지만 WARC의 Revisit 유형은 애플리케이션 단

계에서 중복을 필터링하여 Revisit 유형으로 웹 콘텐츠를 논리적으로 연결하는 구조이다. 그렇기 때문에 이미 저장된 웹 기록물의 Response 유형과 연결하기 위해서는 Revisit 유형 구조에 Response 유형을 변형해야 한다.

[표 4] WARC Revisit 유형의 필드들

구분	설명
WARC/0.17	- WARC 포맷 버전
WARC-Type	- WARC 헤더 유형
WARC-Target-URI	- 수집되는 웹 콘텐츠들의 URI
WARC-Profile	**- WARC 필드 처리 방법을 결정하기 위한 프로파일**
WARC-Date	- 저장된 날짜
WARC-Payload-Digest	- 바이너리에 대한 메시지 다이제스트 값
WARC-Refers-To	**- 참조할 웹 기록물의 UUID 값**
WARC-Record-ID	- 수집되는 웹 콘텐츠의 URI에 대한 UUID
Content-Type	- 수집되는 웹 콘텐츠의 MIME 유형
Content-Length	- Content Header와 Payload를 전체 바이트 크기

Response 유형을 Revisit 유형으로 변형시키는 것은 다음과 같이 세 가지로 구분할 수 있다.

(1) 추가 : Revisit 유형에 추가하는 메타데이터
(2) 수정 : Revisit 유형에 맞게 Response 유형의 메타데이터를 변경
(3) 삭제 : Revisit 유형에 필요 없는 부분으로 삭제되는 필드

Response 유형은 웹 콘텐츠가 저장된 형태로 Revisit 유형으로 변형하기 위해서는 위와 같은 세 가지 조건을 만족해야 한다. [표 5]는 이들 조건을 만족하기 위해서 Revisit 유형의 메타데이터를 변형하기 위한 세부

적인 설명이다. 표에서는 변경해야 할 부분을 세 가지 메타데이터 영역으로 구분하였다. 각 영역은 WARC 메타데이터를 저장하는 WARC Header와 HTTP 프로토콜을 저장하는 Content Header 그리고 웹 콘텐츠의 바이너리 데이터를 저장하는 Payload로 구분된다. 이들 예는 그림 2에서 확인할 수 있다.

Revisit 유형에 추가되어야 할 부분은 이미 수집된 웹 콘텐츠의 UUID를 Reference-ID로 추가하여 논리적으로 연결하는 것이다. 그리고 수정 필드는 Response 유형의 WARC Header를 Revisit 유형으로 변경하고, Content Header의 응답코드를 304로 변경한다. 마지막으로 삭제 필드는 Content Header의 Last-Modified와 웹 콘텐츠가 저장된 Payload의 바이너리 데이터를 삭제하고, 최초로 수집된 웹 콘텐츠를 논리적으로 연결하여 중복을 제거할 수 있게 된다.

이와 같이 방식은 웹 아카이빙 WARC 파일 내에 있는 대량의 의미 없는 웹 콘텐츠들을 줄임으로써 웹 아카이빙 WARC 파일 즉 웹 기록물의 질을 높일 수 있다.

[표 5] 중복 제거를 위한 WARC 메타데이터 변경

구분	메타데이터 영역	설명
추가 필드	WARC Header	- 원본 UUID를 참조한 Reference-ID 추가
	Content Header	
	Payload	
수정 필드	WARC Header	- Response를 Revisit으로 변경
	Content Header	- HTTP 응답코드를 304로 변경
	Payload	
삭제 필드	WARC Header	
	Content Header	- Last-Modified 필드 삭제
	Payload	- Payload 부분 삭제

5. 결론

표면 웹 기록물 수집은 빠르게 생성되고 소멸되는 웹의 특수성을 고려하여 웹 사이트의 전체 자원을 주기적으로 수집하는 방식을 고수하고 있다. 그렇기 때문에 웹 콘텐츠들의 중복 문제가 불가피하게 발생한다. 특히, 웹 서버 환경이 진화하면서 정적 웹 콘텐츠의 중복보다는 웹 프로그램에 의해 자동 생성되는 동적 웹 콘텐츠들의 중복이 많이 발생한다. 또한 웹 기록물을 원격수집하는 것은 웹이라는 복잡한 환경에서 자동화된 도구인 웹 크롤러를 사용하는 것이기 때문에 다양한 문제가 발생할 가능성이 있다.

본 논문에서는 공공기관 수집 사례를 통해서 발생한 중복 문제를 기술 측면에서 분석 및 유형화 하였다. 각 유형 중 의미 없는 웹 콘텐츠의 중복은 다양한 URI에서 무작위로 발생한다는 것을 알 수 있었다. 그렇기 때문에 이들을 필터링 하는데 어려움이 발생한다. 이를 해결하기 위해서 URI 중심의 필터링 방식에 웹 콘텐츠 내용 위주의 필터링 방안을 추가하였다. 또한 WARC 포맷에 이미 중복되어 존재하는 웹 콘텐츠를 필터링하기 위해서 기존의 Reponse 유형을 Revisit 유형으로 레핑할 때 요구되는 메타데이터를 도출하였다. 이들 2가지 방안의 기술 적용을 통해 의미 없는 웹 콘텐츠들에 대한 중복 수집을 효과적으로 필터링할 수 있다고 본다. 나아가, 의미 없는 웹 콘텐츠의 수집을 사전에 차단함으로서 원격수집의 품질을 높이고, 저장 공간을 절감하는데도 기여할 수 있을 것이다.

[참고문헌]

이연수·남성운·박성배, 「공공기관 웹 사이트 기록물 수집 사례분석」, 2010년도 한국멀티미디어학회 추계학술발표논문집, 2010.

Arvidson, A, "Kulturarw3." Proc, Preserving the Present for the Future, 2001.

Brewington, B.E. and Cybenko, G, "How dynamic is the web?", *Computer Networks* 33(1), 2000.

Gomes, D. and Freitas, S. and Silva, M, "Design and selection criteria for a national web archive", Research and Advanced Technology for Digital Libraries, 2006.

Hodges, D. and Lunau, C.D, "The National Library of Canada's digital library initiatives", *Library hi tech* 17(2), 1999.

ISO, "Information and documentation—WARC fileformat(ISO 28500: 2009)", [citied 2011.11.07].
<http://www.iso.org/iso/catalogue_detail.htm?csnumber=44717>

Koerbin, P, "The PANDORA Digital Archiving System (PANDAS) and managing Web archiving in Australia: A case study", 4th International Web Archiving Workshop, Bath (UK), September, 2004.

K. Sigurosson, "Incremental crawling with Heritrix", In Proceedings of the 5th International Web Archiving Workshop, 2005.

Leach, P. Mealling, M. and Salz, R, "A Universally Unique IDentifier (UUID) URN Namespace", RFC 4122, 2005.

Mohr, G., Kimpton, M., Stack, M., & Ranitovic, I, "Introduction to Heritrix", Paper presented at the 4th International Web Archiving Workshop. Bath (UK), September 16, 2004.

Masanès, J, "Web archiving methods and approaches: A comparative study",

Library trends 54(1), 2006.
Phillips, M. E, "What should we preserve? The question for heritage libraries in a digital world", *Library trends*, 54(1), 2006.
Rauber, A. and Aschenbrenner, A. and Witvoet, O. and Bruckner, R. M. and Kaiser, M, "Uncovering information hidden in Web archives", *D-Lib magazine* 8(12), 2002.
Sigurðsson, K, "Managing duplicates across sequential crawls", proceedings of the 6 International Web Archiving Workshop, 2006.

ABSTRACT

A study on the enhanced filtering method of the deduplication for bulk harvest of web records

Lee, Yeon-Soo·Nam, Sung-un·Yoon, Dai-Hyun

As the network and electronic devices have been developed rapidly, the influences the web exerts on our daily lives have been increasing. Information created on the web has been playing more and more essential role as the important records which reflect each era. So there is a strong demand to archive information on the web by a standardized method. One of the methods is the snapshot strategy, which is crawling the web contents periodically using automatic software. But there are two problems in this strategy. First, it can harvest the same and duplicate contents and it is also possible that meaningless and useless contents can be crawled due to complex IT skills implemented on the web.

In this paper, we will categorize the problems which can emerge when crawling web contents using snapshot strategy and present the possible solutions to settle the problems through the technical aspects by crawling the web contents in the public institutions.

Key words: web records, web contents, web crawling, deduplication