반복적 웹 검색을 제거한 다중 웹정보 뷰어

이정수, 이상호
한국산업기술대학교 컴퓨터공학과
e-mail:snp0405@naver.com

Multiple Web-Information Viewer removing repetitive web searching

Jung-Soo Lee, Sang-Ho Lee
Dept of Computer Engineering, Korea Polytechnic University

요 약

인터넷 이용자 급증으로 정보들은 무한히 생산되고 사방에 산재되어 가고 있다. 이로 인해 정보들을 탐색하는 시간은 계속 증가하고 있다. 특히 공지사항이나 날씨처럼 반복적으로 갱신되는 정보들을 얻기 위해 사람들은 동일한 정보를 주기적으로 검색하고 있으며 이에 따른 불필요한 트래픽 유발 및 검색시간이 낭비되고 있는 실정이다. 본 논문은 동일한 정보를 주기적으로 검색함으로써 야기되는 문제점을 서술하고 이를 해결하기 위해 다수의 웹상에서 각종 정보들만을 추출하여 하나의 웹페이지 내에 배치하는 웹 컴포넌트를 설계 및 구현한다. 이 시스템을 사용한다면 사용자는 단순히 하나의 웹페이지를 클릭함으로써 다수의 웹상에 저장된 정보들을 웹서핑 없이 얻을 수 있기 때문에 정보검색 시간을 크게 단축시킬 수 있다. 이 시스템을 구현하기 위해 크로스 도메인상의 웹문서에서 정보를 추출하고 조작하는 것을 금지하는 웹 표준 정책인 동일출처정책을 우회할 수 있는 방법을 서술하였으며 이 정책을 회피함으로써 파생되는 문제점과 해결방안을 서술하였다. 마지막으로 현존하는 관련 시스템들과 비교하여 우수성을 보인다.

1. 서론

많은 사람들이 날씨, 식료품 할인, 주식 등등의 정보를 획득하기 위해 주기적으로 동일한 사이트에 접속하며 상당한 시간을 소요하고 있다. 또 이러한 정보들은 하이퍼링크를 클릭하고 여러 홈페이지를 이동한 끝에 찾아낼 수 있으며 이러한 검색방법은 불필요한 네트워크 트래픽과 시간을 낭비한다. 각각의 웹페이지가 HTML 문서 1개와 5개의 이미지 객체들을 가지고 있다고 가정하고 연결설정과 HTML 문서를 받는데 걸리는 시간을 RTT (Round Trip Time), 웹 페이지 내 객체들을 받는 시간은 TR(Transmission Time)이라고 했을 때 파이프라인을 이용하여 접속한다고 가정했을 때 1개의 웹페이지에 접근하는 시간은 RTT+5TR만큼의 시간이 걸린다. [1]

하지만 위에서 설명한 바와 같이 하이퍼링크를 클릭하고 이동한 횟수를 N이라 한다면 원하는 정보 하나를 얻기 위해서는 N*(RTT+5TR)로써 O(n)의 시간이 걸린다. 이것은 정보를 얻기 위해 클릭해야하는 하이퍼링크의 수가 많아짐에 따라 검색시간과 트래픽은 크게 늘어난다는 것을 의미한다.

또한 현재 검색 시스템은 검색어에 대한 가능성 있는 모든 정보들을 보여준다. 때문에 내가 원하지 않는 정보들까지 봐야 된다는 단점이 존재한다. 예를 들어 NAVER 검색엔진에 날씨로 검색을 했을 시 날씨 정보는 한 페이지 내에서 10%도 차지하지 않고 있으며 실시간 검색어, 어학사전, 광고, 뉴스들이 90% 이상을 차지하고 있는 것을 확인할 수 있다.

위에서 제시한 문제점들을 해결하기 위해선 다수의 웹상에서 원하는 정보들을 추출하여 특정 뷰어 페이지 내에 배치한다. 사용자는 단순히 이 뷰어 페이지를 열어서 매일 검색해야 하는 정보들을 모아서 본다면 검색시간 단축과 불필요한 네트워크 트래픽을 감소시킬 수 있다.

이 시스템을 구현하기 위해서는 타도메인에 위치한 웹페이지에서 특정정보가 존재하는 범위좌표를 추출해야 하며 이 정보를 사용하여 뷰어 페이지 내에 원하는 정보만을 배치한다. 이 방법은 웹 표준 정책인 동일출처정책(Same Origin Policy)에 의해서 금지되어 있기 때문에 이 정책을 회피할 수 있는 연구가 필요하다.

본 논문의 구성은 다음과 같다. 2장에서 웹사이트들의 특정한 정보들을 한 페이지에 모아서 보기 위해 개발된 기존 시스템들을 소개하고 이 시스템들이 가지고 있는 문제점을 설명한다. 3장에서는 2장에서 언급한 기존 시스템들의 문제점을 해결하기 위해 동일출저정책을 회피하는 방법과 회피함으로써 파생되는 문제점 및 해결방안을 서술한다. 4장에서는 기존의 시스템과 성능평가를 하여 우수성을 보이고 마지막으로 5장에서는 결론을 맺는다.

2. 관련 시스템

웹페이지 내 정보들을 모아서 보기 위한 시스템 중 널리 알려진 것은 3만명의 유저를 가지고 있는 “ichrome”이라는 크롬애플리케이션이 있다. 하지만 동일출처정책의 한계에 벗어나지 못하여 특정 부분만을 보여주지 못하고 전체 페이지만을 보여주는 방법을 채택하고 있기 때문에 사용자는 다시 페이지 내 특정 정보가 위치한 부분으로 스크롤

바를 사용해야 한다는 단점이 있다. 이 외에 6만명의 유저가 이용중인 "Screen capture" 크롬 애플리케이션은 동일출처정책을 회피하기 위해 화면캡처방식을 사용했으며 "ichrome"에서 특정 부분을 보여줄 수 없는 문제점을 해결하였다. 하지만 이 방법은 웹상에서 동영상 또는 플래시는 재생이 불가능하며 최신의 정보가 업데이트 되었을 시 다시 화면을 캡처를 해야 한다는 번거로움이 존재한다.

3. 동일출처정책과 해결방안

동일출처정책이란 보안을 위해 클라이언트 스크립트는 다른 도메인(Cross Domain)에 존재하는 문서와 상호작용할 수 없으며 스크립트는 오로지 자신이 포함된 문서와 출처가 동일한 문서나 창의 프로퍼티만을 읽을 수 있게 한 웹 표준을 말한다. 본론2에서 제시한 문제점을 해결하기 위해서는 타 도메인에 위치한 사이트에서 정보를 추출해야 하나 동일출처정책에 의해 금지되어 있다. 따라서 이 정책을 우회할 수 있는 연구가 필요하다. 3.1과 3.2에서는 동일출처정책을 우회하는 방안을 제시하며 이 과정에서 생겨난 문제점과 해결방안을 3.3과 3.4에 서술하였다.

3.1 타도메인의 웹페이지에서 정보가 위치하는 좌표추출

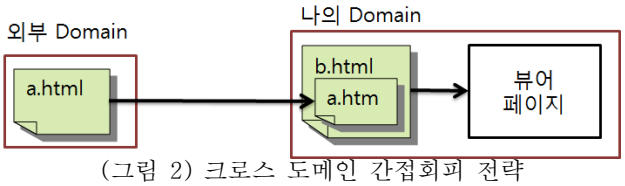
웹상에서 사용자가 원하는 정보만을 추출하기 위해서는 타 도메인에 위치한 웹페이지에 접근 후 정보를 받아와야 한다. 이 방법은 사용자는 Iframe을 통해 웹서핑을 하다가 원하는 정보를 찾을시 이것이 존재하는 부분을 드래그하여 서버로 특정정보가 존재하는 좌표정보를 전송하면 된다. 하지만 이 방법을 구현하기 위해서는 드래그 이벤트를 크로스도메인 상의 웹페이지에 설정해둘 필요가 있다.

(그림 1) 드래그이벤트를 담은 투명 레이어와 범위 좌표

이 방법은 동일출처정책에 의해 금지되어 있으므로 사용자가 드래그 이벤트를 수행하기 직전 (그림1)과 같이 Document 전체에 투명한 레이어를 생성하고 이곳에 드래그 이벤트를 설정한다면 사용자의 드래그를 받아 좌표정보를 추출해낼 수 있다.

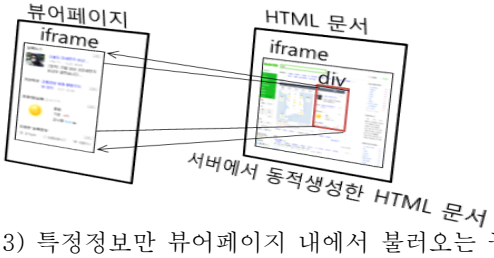
3.2 추출해낸 좌표정보를 토대로 뷰어 페이지 내에 정보들을 배치하기

3.1절에서 추출해낸 좌표 정보를 바탕으로 타도메인에 위치한 웹페이지의 특정 범위만을 보여주는 방법은 동일출처정책으로 인해 금지되어 있다. 위 문제를 해결하기 위해 화면을 캡처하여 이미지 파일을 자신의 도메인에 저장하여 보여주는 방법이 있지만 이 방법은 본론2에서 제시한 것처럼 많은 문제점을 만들 수 있다.



(그림 2) 크로스 도메인 간접회피 전략

이를 해결하기 위해 (그림2)와 같이 서블릿에서 HTML 문서를 동적 생성하고 이 문서에 내가 원하는 정보가 존재하는 웹페이지를 Iframe을 생성한다.[2] 이때 Iframe의 크기는 웹페이지의 전체 분량을 담을 정도의 크기여야 한다.



(그림 3) 특정정보만 뷰어페이지 내에서 불러오는 구조

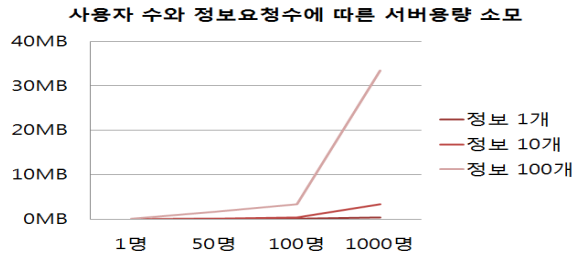
다음으로 Iframe 상위에 Div 태그를 생성하여 3.1절에서 추출해낸 좌표에 위치시킨다. 이렇게 만든 HTML 문서는 내 도메인 상에 위치하면서도 외부 도메인에 위치한 웹사이트의 특정 범위만을 표현하게 된다. (그림3)과 같이 이 HTML 파일은 Cross Domain이 아니므로 뷰어 페이지에서는 만든 Iframe에 이 HTML 파일을 지정하여 불러올 수 있다.

(그림 4) 정보 5개를 담은 하나의 웹페이지

(그림4)는 3.1, 3.2에서 제안한 방법으로 만든 뷰어페이지를 나타낸다. 이 뷰어페이지는 NAVER 주요뉴스, DAUM 주요뉴스를 비롯한 URL이 각각 다른 사이트에서 추출한 5가지 정보를 가지고 있으며 본론2에서 언급한 기존 시스템들의 단점을 해결한 것을 볼 수 있다. 사용자는 위 5가지 정보를 얻기 위해 웹서핑할 필요 없이 단순히 위 페이지만을 불러오면 된다.

3.3 서블릿에서 동적생성되는 HTML 파일 누적으로 인한 과부하 문제점

3.1과 3.2에서 저자가 제안한 방법은 정보의 개수에 따라 새로운 웹페이지를 만들어 내 도메인 내에 저장해야 하는 문제점을 가진다. 정보 1개를 담고 있는 페이지 1개가 생성될 때마다 평균적으로 약 350Byte의 용량을 차지한다. 수많은 이용자가 다수의 정보들을 뷰어 페이지 내에 저장하고 불러온다면 서버의 용량은 머지않아 초과될 것이고 의도치 않은 DOS 상황이 발생할 것이다.



(그림 5) 동적생성되는 HTML파일로 인한 서버 용량 소모

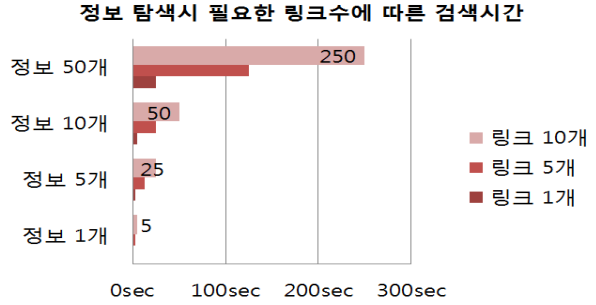
(그림5)은 사용자의 수와 이 사용자들이 요청하는 뷰어 페이지 내부에 저장된 정보의 수를 기준으로 서버에서 동적 생성되는 데이터 크기를 나타낸다. 이것은 웹페이지를 교체해주는 정책을 사용하지 않는다면 서버의 부하는 기하급수적으로 증가하게 된다는 것을 의미한다.

3.4 웹페이지(정보) 교체 전략

서버의 용량은 한정되어 있고 인터넷에 존재하는 모든 웹페이지를 저장할 수 없으므로 이를 해결하기 위해 생성된 웹페이지(정보) 교체 전략이 필요하다. LFU, LRU, Clock과 같이 검증된 페이지 교체 알고리즘이 있다. 하지만 인터넷에 존재하는 정보들은 갱신되는 시간이 각각 다르고 이에 따라 참조되는 시간들도 다르다. 실제 저자가 인터넷 이용자들을 조사해본 결과 하루를 기준으로 날씨나 공지사항은 평균 1번, 뉴스정보는 평균 10번, 쇼핑정보는 평균 3번 참조하는 사실을 알 수 있었다. 따라서 참조된 빈도수와 참조된 시간뿐만 아니라 참조된 시간 간격을 동시에 고려하여 동적 생성된 웹페이지(정보)들을 교체해 줄 필요가 있다. 서버는 참조된 시간과 참조 횟수를 테이블에 저장하고 유지하며 이를 기반으로 Clock 알고리즘을 사용하여 교체대상을 고른다. 그 후 참조된 주기를 조사하여 웹페이지가 주기적으로 호출되었던 것이라면 교체순위를 뒤로 옮긴다. 제약조건으로 동적 생성된 웹페이지는 일정 기간 동안 Clock 알고리즘에 의해 교체되지 않고 참조된 시간을 테이블에 기록 유지하고 패턴을 분석함으로써 주기적으로 호출되는 웹페이지인지 알아내야 한다. 위와 같은 방법으로 정보 수요를 예측한다면 다른 교체 전략에 비해 서버의 성능을 높일 수 있으며 DOS 상황을 방지할 수 있다.

4. 성능평가

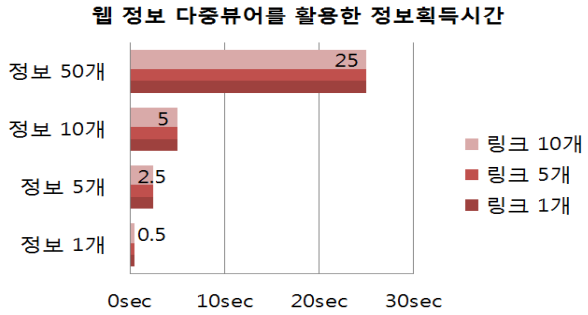
본 장에서는 기존의 정보 검색 방법과 저자가 제시한 시스템을 사용하여 정보를 검색하는 시간을 측정한 결과를 서술하였다. 웹페이지를 불러오는 시간은 네트워크 트래픽의 상황에 따라 다르므로 500ms라 가정하였다.



(그림 6) 기존의 시스템의 정보검색시간

(그림6)에는 기존의 여러 정보들을 검색하는 방법을 사용했을 때의 시간을 나타낸다. 검색해야 하는 정보의 수와 그 정보를 얻기 위해 클릭해야하는 하이퍼링크의 수에 따라서 검색시간이 급속도로 증가하는 것을 보여준다.

반면 (그림7)는 저자가 제시한 시스템을 사용했을 때의 검색시간을 나타낸다. 정보를 얻기 위해 하이퍼링크를 타고 이동할 필요가 없으므로 오로지 검색해야 하는 정보들의 수에 따라서 일정한 시간이 걸린다.



(그림 7) 다중 웹정보 뷰어의 정보검색시간

기존의 검색 방법과 비교했을 때 정보검색시 이동해야할 하이퍼 링크 수가 10개일 때 최대 1/10까지 줄어들음을 확인할 수 있다. 이런 정보검색 시간의 우수성 외에도 특정 뷰어페이지 하나에 사용자가 원하는 정보들만 모아놓을 수 있기 때문에 사용자는 원치 않은 광고나 다른 정보들을 볼 필요가 없으므로 웹페이지 활용률은 100%가 된다.

4. 결론

본 논문에서는 반복적으로 갱신되는 특정 정보들에 대해 주기적으로 검색함으로써 소요되는 시간과 네트워크 트래픽 문제점, 검색 결과 시스템의 페이지 활용률 저하라는 문제점을 제시했으며 이를 해결하기 위해 한 웹페이지 내에 여러 웹 정보들을 저장하는 방법을 제시하였다. 본론에서는 관련 시스템들을 소개하고 이들의 공통적인 문제점을 만든 웹 표준 동일출처정책에 대해 소개하였다. 동일출처정책을 우회하기 위해 서버는 동적으로 웹페이지를 생성해야 하는데 이로 인한 서버 과부하 문제점과 페이지 교체전략을 통한 해결방안을 제시하였다. 마지막으로 본 논문이 제시한 방법으로 개발한 시스템과 기존의 시스템의 정보검색시간을 비교함으로 우수성을 보였다.

본 논문에서 구현된 다중 웹 정보 뷰어 시스템은 사용자들의 정보검색시간을 최소화하고 불필요한 웹 검색을 제거함으로 네트워크 트래픽을 효율적으로 줄일 수 있다. 또한 웹페이지 활용률을 높여주며 사용자들의 웹 정보 관리가 용이하다. 본 연구는 인터넷 정보 검색 방법의 새로운 방향을 제시할 것이라 생각된다.

참고문헌

[1.] James F. Kurose and Keith W. Ross, Computer Networking A Top-Down Approach Featuring the Internet 5th Ed, PEARSON, March 2009

[2.] Bryan Basham and Kathy Sierra, Head First Servlet & JSP 2nd Ed, ORELLY, March 2008