WebBase 기반 웹 아카이브 시스템의 설계

이민희*, 이무훈*, 장창복*, 김동혁*, 고병오**, 최의인*
*한남대학교 컴퓨터공학과
**공주교육대학교 컴퓨터교육과
e-mail:mhl@dblab.hannam.ac.kr

The Design of Web Archive System on the WebBase

Min-Hee Lee*, Moo-Hoon Lee*, Chang-Bok Jang*, Dong-Hyuk Kim*, Byoung-Oh Goh**, Eui-In Choi*
*Dept of Computer Engineering, Han-Nam University
**Dept of Computer Education, Gongju National University of Education

요 약

웹의 성장으로 사용자는 언제 어디서든지 유용한 정보의 이용이 가능해졌다. 웹이 광범위하게 사용됨에 따라 정보를 획득하기 위해 대다수의 사용자들이 웹을 의존하고 있다. 그러나 웹상의 모든 정보는 정보가 저장되어 있는 서버의 관리자들에 의해 계속적으로 갱신 또는 삭제되어 가고 있어 기존의 정보들은 그것의 중요성 여부와 관계없이 대다수의 정보가 소멸되고 있다. 따라서 오랜 기간에 거쳐 생성된 웹상의 중요 데이터(importance data)들을 효율적으로 활용하기 위한 웹 아카이브(archive) 시스템이 연구되었다. 그러나 현재 존재하는 웹 아카이브 시스템은 갱신되기 이전의 데이터를 다루기 위한 체계적인 처리기법을 제시하지 못하고, 수집된 데이터들에 대한 연관관계를 저장하지 못하여 데이터 관리에 있어 비효율적이라는 문제점을 가지고 있다.

이에 따라 본 논문에서는 웹으로부터 다운로드한 데이터를 레포지토리(repository)에 효율적으로 저장하기 위해 설계된 대표적인 WebBase를 기반으로 하여 갱신되기 이전의 모든 정보들에 대한 내용을 히스토리(history) 저장소내에 저장하여 정보를 효율적으로 활용할 수 있는 웹 아카이브 시스템의 구조를 제안한다.

1. 서론

지난 10년간 웹이 엄청난 속도로 성장함에 따라 사용자는 시공간의 제약없이 유용한 정보의 이용이 가능해졌다. 또한 웹을 이용하는 사용자의 요구 또한 점차적으로 복잡해지고 사용자의 수도 급격하게 증가하고 있다. 웹이 광범위하게 사용됨에 따라 정보를 획득하기 위한 웹의 의존도도 높아지게 되었다. 따라서 웹의 효율적인 정보관리의 필요성이 크게 증대되고 있다. 그러나 현재 웹상의 정보관리는 최신의 정보에 대한 획득과 관리가 효율적으로 이용되고 있기는 하지만 최신 정보로 갱신되기 이전의 정보에 대한 관리가 미비하다는 문제점이 발생하게 되었다. 즉, 현재 웹상의 모든 정보는 정보가 저장되어 있는 서버의 관리자들에 의해 계속적으로 갱신 또는 삭제되어 가고 있어 기존의 정보들은 그것의 중요성 여부와 관계없이 계속적으로 소멸되고 있다 [1, 2, 3]. 따라서 오랜 기간에 거쳐 생성된 웹상의 중요 데이터(importance data)들을 효율적으로 활용하기 위하여 웹상의 정보들의 갱신되기 이전의 모든 정보들에 대한 내용을 히스토리(history) 저장소 내에 저장하여 정보를 효율적으로 활용할 수 있는 웹 아카이브 시스템에 대한 연구가 필요하다. 현재 웹 아카이브 시스템에 관하여 다양한 연구가 진행 중이기는 하나, 히스토리 데이터를 다루기 위한 체계적인 처리기법을 제시하지 못하고, 수집된 데이터들에

본 연구는 한국과학재단 지역협력연구사업(R12-2003-004-03002-0)지원으로 수행되었음.

대한 연관관계를 저장하지 못하여 데이터 관리에 있어 비효율적이라는 문제점을 가지고 있다[3].

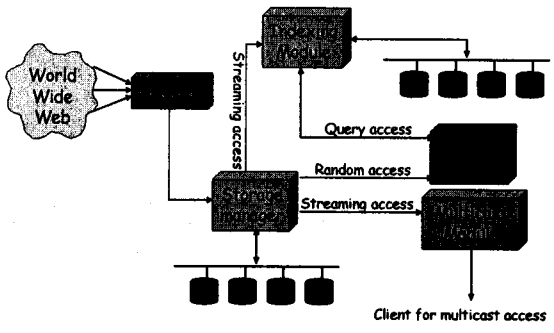
따라서 본 논문은 기존 웹 아카이브 시스템의 문제점을 해결하고자, 히스토리 데이터를 저장할 수 있는 웹 저장 시스템을 설계하여 웹상의 정보들의 변화를 추적하고 웹 문서의 다양한 버전(version)을 간결하게 저장하는 시스템에 관한 연구이다.

논문의 구성은 2장에서는 관련연구로서 본 연구의 기반이 되는 시스템과 기존 웹 아카이브 시스템에 대하여 알아보고, 3장에서는 제안한 웹 아카이브 시스템의 구조와 각 모듈별 주요 처리과정을 설명한 뒤 마지막으로 4장에서 결론 및 향후 연구내용을 기술한다.

2. 관련연구

2.1 WebBase

WebBase는 스탠포드 대학에 개발 중인 프로젝트로, 웹으로부터 다운로드된 데이터가 저장되어 있는 레포지토리(repository)를 효율적으로 관리하기 위하여 설계된 대표적인 웹 검색 엔진이다. 수집된 웹 페이지의 거대한 집합에 대하여 스토리지(storage), 인덱싱(Indexing), 쿼리(querying) 등 다양한 관점으로 연구가 진행 중이다. WebBase의 프로토타입(prototype)은 대략 4천만 개의 웹 페이지들의 집합을 가지고 서로 다른 스토리지, 인덱싱, 데이터 마이닝(data mining) 기술을 연구하는 시험대로서 사용중이다. 프로토타입의 초기 버전은 Google 검색 엔진의 back-end 스토리지 시스템으로 사용하였고 새로운 프로토타입은 다양한 스토리지 컴퓨터를 거쳐서 병렬처리를 제공하며 다양한 애플리케이션들을 제공한다[2]. 현재 WebBase가 제안하는 특성들과 컴포넌트들이 모두 구현되지는 않았지만, 대부분 중요



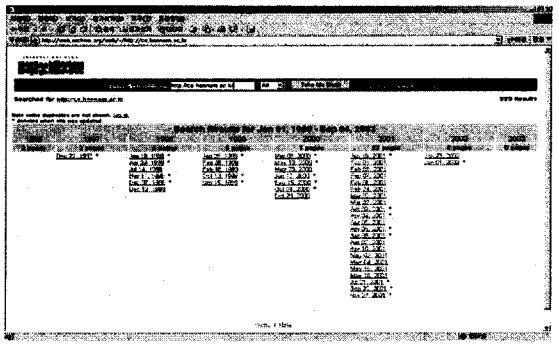
[그림 1] WebBase의 구조

기능들과 서비스들은 이미 적용중이다. WebBase의 구조는 [그림 1]과 같다.

WebBase는 데이터 접근을 최적화하고, 분산된 데이터베이스 관리 시스템 내에 모든 메타데이터를 저장하는 기법을 제시하기는 하나 갱신되기 이전의 데이터에 대한 관리기법을 제시하지 못하는 단점을 가지고 있다[2, 3].

2.2 Internet Archive WayBack Machine

비영리 집단인 Internet Archive와 Alexa Internet이 공동으로 개발한 시스템에 저장된 오래된 웹 페이지에 대한 접근을 시도한 것이다. WayBack Machine은 Internet Archive의 웹 저장 공간에 1억 페이지 이상의 웹 페이지들을 저장하여 웹 페이지의 다양한 버전을 관리, 저장, 압축, 다운로드하기 위한 정책들을 연구하고 있다. 이 시스템은 실제적으로 2001년 U.C.Berkeley의 Bancroft Library에서 이용하고 있다[4, 5, 6]. 다음의 [그림 2]는 WayBack Machine의 실행화면을 나타낸다.



[그림 2] WayBack Machine

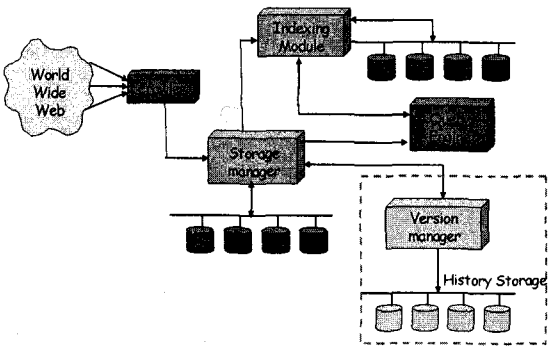
WayBack Machine은 URL 검색을 통하여 수집된 데이터를 디지털 형식으로 저장하고, 선택된 컨텐츠에 대하여 오랜 기간의 보존이 이루어지기는 하나 웹 히스토리 데이터를 다루기 위한 체계적인 처리기법을 제시하지 못하고, 수집된 데이터에 대한 연관관계를 저장하지 않는다는 단점을 가지고 있다.

3. 웹 아카이브 시스템의 구조 및 주요 처리 모듈

3.1 웹 아카이브 시스템의 구조

본 논문에서 제안하는 웹 아카이브 시스템은 WebBase에 기반을 두고 있다. Crawler를 통해 수집된 데이터들이 Storage Manager에 의해 저장소에

갱신되면서, 저장소 내에 저장되어 있던 기존 데이터가 삭제되면 삭제된 데이터, 즉 히스토리 데이터를 Version Manager를 통해 History Stroage에 저장, 관리할 수 있도록 하였다. 이때 Version Manager는 여러 개의 처리모듈을 두어 버저닝(versioning)을 수행함으로서 수집된 데이터의 체계적인 관리가 가능하도록 하였다. 또한 Version Manager를 통하여 처리된 히스토리 데이터는 History Storage에 저장하게 되는데, History Storage의 저장기법은 WebBase의 저장기법과 동일한 방법으로 저장한다. [그림 3]는 본 논문에서 제시하는 웹 아카이브 시스템의 구조이다.



[그림 3] WebBase 기반 웹 아카이브 시스템의 구조

3.2 웹 아카이브 시스템의 주요 처리 모듈

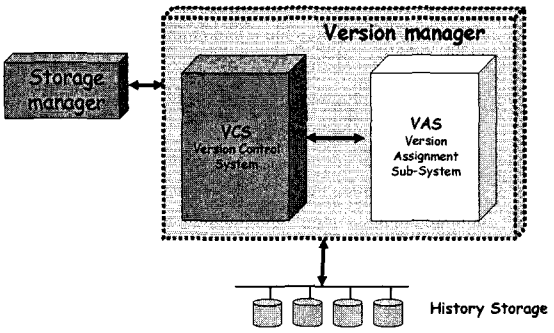
본 논문에서 제안한 시스템의 주요 처리 모듈은 Version Manager이다. Version Manager는 기존 파일 시스템의 버저닝 구조인 Moraine 시스템을 응용하여 웹에 접목한 것으로, VCS(Version Control System)와 VAS(Version Assignment System)로 두 가지 핵심 모듈로 구성하였다. Version Manager에 대한 세부구조는 [그림 4]과 같다.

VCS는 History Storage의 각 노드에 대한 최신 정보리스트를 이용하여 기존 WebBase의 Storage Manager에서 삭제된 데이터와 History Storage내에 존재하는 데이터를 비교하고, 비교된 데이터는 버전 할당을 위해 VAS로 전송한다. 이 때, 전송된 데이터는 최신정보리스트에 대한 식별자 존재 여부에 따라 New 데이터와 Old 데이터로 분류한다.

또한 VAS에서 처리된 데이터에 대하여 최신정보리스트의 갱신을 수행하고 History Storage로의 저장을 수행한다.

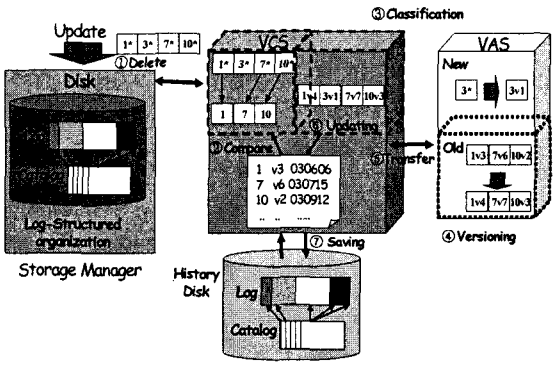
VCS의 최신정보리스트에는 각 노드의 데이터에 대한 식별자, 데이터의 버전정보, 타임스탬프(Timestamp)가 기록되어 있다.

VAS는 VCS에서 분류된 데이터에 대하여 각기 버전할당을 수행하고, 할당된 데이터를 VCS로 전송한다.



[그림 4] Version Manager의 세부구조

History Storage는 히스토리 데이터를 저장하는 데이터 저장소로서 WebBase의 Storage와 같이 여러 개의 노드로 분산하여 데이터 저장을 최적화하고, 각 노드별 저장은 Log-Structured 정책을 이용한다. 노드의 구성은 Log와 Catalog로 구성되어 있으며 Log는 히스토리 데이터가 실제적으로 저장되는 부분이고, Catalog는 히스토리 데이터의 식별자, Log내에 데이터의 물리적 위치 정보를 나타내는 포인터, 데이터의 크기와 같은 데이터의 메타 정보와 함께 데이터의 버전정보를 포함한다. 히스토리 데이터에 대한 처리 시나리오는 [그림 5]와 같다.



[그림 5] 히스토리 데이터의 처리절차

Storage manager를 통해 삭제된 데이터는 Version Manager의 VCS로 전송된다. VCS는 각 노

드별 최신정보리스트를 유지하고 있으며, 이 리스트의 정보를 이용하여 데이터 분류 후 VAS로 전송한다. VAS는 분류된 데이터에 따라 각각의 버전 정보를 할당하고, 할당된 데이터는 VCS로 전송한다. 새로운 버전을 할당받은 데이터는 최신정보리스트를 갱신한 후 History Storage내 노드에 저장된다.

히스토리 데이터의 처리 절차는 다음과 같다.

단계 1 : Send Deleted Data in Storage Manager To VCS of Version Manager.
단계 2 : Compare Inserted Data of VCS with Data of Lastest Data List
단계 3 : Classify New Data and Old Data in VAS
단계 4 : Versioning Data
단계 5 : Transfer Versioned Data to VCS
단계 6 : Update Lastest Data List
단계 7 : Save Data in History Storage

4. 결론 및 향후 과제

본 논문에서는 웹 검색 시스템인 WebBase를 기반으로 히스토리 데이터를 저장할 수 있는 웹 아카이브 시스템을 제안하였다. 이 시스템을 통하여 히스토리 데이터를 보다 체계적으로 관리할 수 있고, 수집된 데이터들 간의 버전 정보를 할당하여 데이터들 간의 연관관계를 생성하였다.

향후 연구과제로는 히스토리 데이터에 대한 다양한 활용을 위하여 효율적인 검색이 가능하도록 하기 위한 인덱싱 기법에 관한 연구가 수행되어야 한다.

참고문헌

[1] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan, Searching the web (invited paper). ACM Transactions on Internet Technology, 1(1), August 2001

[2] Jun Hirai, Sriram Raghavan, Hector Garcia-Molina, and Andreas Paepcke. Webbase: A repository of web pages. In Proceedings of the International World-Wide Web Conference, pages 277 - 293, May 2000.

[3] Joao P. Campos, Versus: a Web Data Repository with Time Support, May 2003, http://www.di.fc.ul.pt/tech-reports.

[4] Internet Archive, http://www.archive.org

[5] Brewster kahle, Archiving the Internet, http://www.uibk.ac.at/sci-org/voeb/texte/kahle.html

[6] Burner, M. Crawling towards eternity: Building an archive of the World Wide Web. Web Techniques Magazine 2, 5 (May 1997).