제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

대규모 웹사이트 갱신 시스템의 설계 및 구현

하동근, 정성주, 박희숙, 조우현
부경대학교 컴퓨터공학과
e-mail:wininfo@korea.com

Design and Implementation of system for Refreshing a Very Large Website

Dong-Keun Ha, Sung-Ju Jung, Hee-Sook Park, Woo-Hyun Cho
Dept of Computer Engineering, Pukyong National University

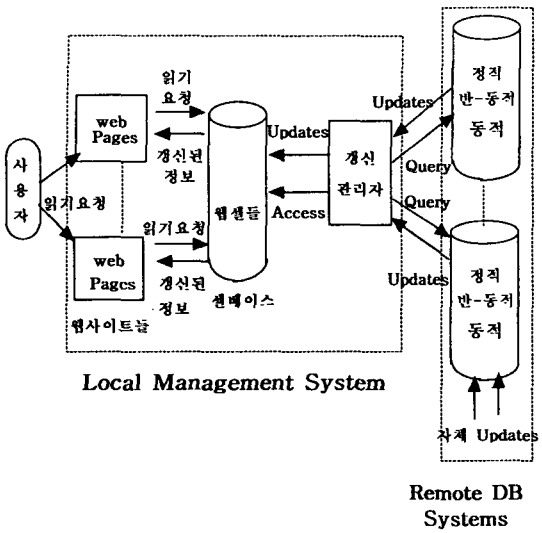
요 약

웹사이트들의 급성장과 더불어 사용자들에게 유용한 정보를 제공하기 위해 대규모 웹사이트의 갱신 문제는 그 신선도 유지를 위해 적절한 시간 내에 요청된 갱신질의어를 수행하기 위한 많은 연구들이 진행되어 왔다. 본 논문에서는 동적, 반동적, 정적 웹사이트의 베이스데이터들을 통합 관리하도록 설계 및 구현을 하였다. 웹사이트의 신선도 유지와 기아상태의 공정성을 유지하기 위해 요청된 갱신질의어 실행시간 할당을 위한 목표 갱신비율을 일정한 스케쥴링 주기마다 재계산하는 스케쥴링 알고리즘을 구현하였다

1. 서론

오늘날의 월드와이드 웹은 그 수요에 있어서 폭발적인 증가와 더불어 무한한 잠재적인 능력을 지닌 네트워크로 성장하였다. 이런 배경을 기반으로 하여 많은 기업들과 개인사업자들은 자신들의 정보 전달과 정보저장을 위한 수단으로 웹사이트 사용을 원하고 있다. 따라서 월드와이드 웹은 오늘날 정보 보급의 중요한 수단이 되어 지고 있으며 그 중요성 또한 날로 증가되고 있는 추세이다. 일반적으로 월드와이드 웹에서 제공되는 웹사이트의 종류는 정적, 반-동적, 동적 웹사이트로 분류할 수 있다. 정적 웹페이지(Static Web page)는 시간이 지나도 그 내용이 변하지 않거나 아주 드물게 변화하는 웹페이지를 말한다. 동적 웹페이지(Dynamic Web page)는 요구된 입력 파라미터들을 사용자가 질의 형식으로 전송하였을 때 실행시간에 CGI 스크립트를 이용하여 동적으로 계산되어진다. 동적 웹페이지의 내용은 입력 파라미터에 따라서 그 내용이 다양하게 변화된다. 반-동적 웹페이지(Semi-Dynamic Web page)는 몇몇 소스데이터베이스로부터 유도된 내용을 가지며, 그 내용은 소스 데이터베이스의 갱신(Updates)에 대해 응답하여 변화한다. 인터넷 사용자들은 월드와이드 웹에서 제공되는 정확한 최신의 정보들을 보기를 원한다. 따라서 최신의 정보를 웹사이트에서 유지하는 것은 중요한 문제가 된다. 최신의 정보를 웹상에서 유지하기 위해서는 웹사이트의 정보를 유도하는 베이스테이블에 대한 신선도를 유지하기 위한 적절한 갱신(Refresh)방법이 요구된다. [1]

본 논문에서는 정적 웹사이트, 반-동적 웹사이트 동적 웹사이트를 통합 구현하고 이들의 특징 및 성능을 비교 분석 할 것이다. 또한 베이스테이블의 신선도 유지 문제를 해결하기 위한 새로운 갱신 알고리즘의 모형을 제시하고 구현한다. 또한 기존의 갱신알고리즘과의 특징을 비교 및 분석을 할 것이다. 본 논문은 다음과 같이 구성된다. 2장에 전체적인 시스템의 구성 모델을 제시하고, 갱신 알고리즘에 사용된 수학적인 공식표현을 기술한다. 3장에서는 통합 구현한 내용과 결과를 기술한다. 마지막으로 4장에서 결론을 기술한다.



(그림 1) 웹사이트 갱신 시스템 구성도

2. 시스템의 구성 모델

2.1 전체 시스템의 구성 요소들

인터넷사용자들의 요구를 만족할 수 있도록 가치있고 유용한 정보를 제공할 수 있는 인터페이스로 설계된 웹사이트라 할지라도 소스데이터베이스의 잦은 갱신(Update)과 관련하여 최신의 데이터를 웹사이트에 유지하는 문제는 어려운 것이다. 그 이유는 웹사이트의 신선도(Freshness)는 소스데이터베이스에서 웹사이트들로 베이스데이터를 일방적으로 밀어 넣기(Pushing) 하는 것으로는 웹사이트의 신선도를 보장할 수 없으며, 신선한 데이터는 소스데이터베이스에 대한 갱신질의어 실행에 의해 웹사이트로 신선한 데이터가 제공(Pulled)되어지기 때문이다.[2] 따라서 본 시스템에서는 위에 언급한 3가지 형태의 웹사이트를 통합 구현하였으며, 그 시스템의 구성도는 (그림 1)과 같다. 그 구성요소들은 다음과 같은 것으로 구성되어 져있다.

① 베이스데이터(Base Data): 베이스데이터는 원격지 시스템 내에 관계형 데이터베이스 안에 저장된 테이블 또는 뷰로 구성한다. 그들의 내용은 독립적인 애플리케이션에 의해 자체 갱신된다.

② 갱신질의어(Refresh Query) : 갱신질의어 집합은 베이스데이터와 관련한 질의어들을 수행함으로써 셀베이스의 정보를 최신의 내용으로 유지한다.

③ Web Cell : 웹셀은 웹페이지의 일부분이며 베이스테이블들로부터 유도된 항목들을 포함한다. 한 개의 셀은 하나 이상의 웹페이지에 포함되어질 수 있다. 각 셀들은 갱신주기에 따라 정적셀, 반-동적셀, 동적셀들 중 하나가 된다. 한 개의 셀은 튜플<C_id, C_data>로 정의한다. C_id는 셀 베이스내의 셀 식별자를 나타내며, C_data는 셀베이스내의 셀의 데이터를 나타낸다.

④ 셀베이스(CellBase) : 셀베이스는 웹셀들의 집합으로서 테이블로 구현한다. 그 구성 컨튜플<I, P , Eq, Rq, r>이다. 여기서 I는 셀의 식별자를 의미하며, P는 갱신주기를 나타내며, Eq는 갱신질의어 실행시간을 Rq는 갱신요청횟수, r은 실제 갱신횟수를 나타낸다. 각 웹사이트들은 웹페이지 셀들의 내용으로 셀베이스의 데이터를 참조하기 때문에 셀베이스는 캐쉬의 역할을 할 수 있다는 의미이다.

⑤ Web Page : 사용자의 읽기 요청에 의해 보여지는 갱신된 최신의 웹셀들의 내용을 보여주기 위한 문서를 말한다.

⑥ 갱신 관리자(Refresh Manager) : 갱신관리자는 자원들의 집합으로 프로세스, 네트워크, 전송매체 등을 포함하는 웹사이트의 갱신 작업에 사용되어진다. 갱신관리자는 웹셀들과 베이스데이터의 매핑, 유지관리, 설정 및 셀베이스내에 구현될 셀들의 집합을 결정하고, 웹사이트의 갱신작업, 갱신질의어 우선순위 알고리즘을 수행해야 할 의무가 있다.[1]

2.2 갱신 스케줄링 알고리즘의 모델 제시 및 구현

웹 페이지들의 신선도를 유지하기 위해서는 갱신요청이 일어났을 때 모든 갱신요청들을 만족할 수 있도록 적절한 방법으로 갱신질의어 집합을 실행할 수 있는 스케줄링 알고리즘이 필요하다. 그러나 베이스데이터의 다양한 갱신패턴(Update Pattern)의 주기차이와 실행시간 때문에 몇몇 갱신요청들은 실행을 실패(Missed)하게 된다.[1][3] 셀들의 갱신요청에 대한 실패된 갱신을 기아상태(Starvation)라 한다. 기아상태는 웹 페이지들의 신선도를 저하시키는 직접적인 원인이 된다. 따라서 본 논문에서는 이런 기아상태를 완화하기 위해 일정한 주기마다 예상갱신비율(ν)을 계산하여 이것을 기준으로 각 셀들의 갱신비율을 비교하여 최소 갱신비율을 갖는 셀을 우선적으로 갱신하여 각 그룹별 공정성(Fairness)이 유지되는 알고리즘을 구현한다. 예상 갱신비율(ν)계산은 셀베이스의 정보를 이용하며 다음과 같이 공식을 적용한다.

$$\nu = \frac{\lambda}{\eta} \qquad (1)$$

식(1)에서 ν는 셀의 예상갱신비율을, λ는 셀의 실제

갱신 카운트를 η는 갱신요청 카운트를 의미한다. 식(1)에서 사용된 η의 계산은 다음과 같은 공식을 적용한다.

$$\eta = \sum_{i=1}^{n} \frac{T}{P_i} \qquad (2)$$

식(2)에서 T(Refresh Manager의 1회 스케줄링의 실행시간 * 스케줄링 횟수)는 정해진 시간을 의미하며 P_i는 각 셀의 갱신주기를 나타낸다. 식(1)에서 사용된 셀의 실제 갱신 카운트 λ의 계산은 다음과 같은 공식들을 적용한다.

$$P_s = \sum_{i=1}^{n} P_i \qquad (3)$$

$$E_t = ((P_s/P_i)/ \sum_{k=1}^{n} \frac{P_s}{P_k}) * T \qquad (4)$$

$$Rc = E_t \,/\, Avg(E_{q_i}) \qquad (5)$$

$$\lambda = Min(\ \eta\ ,\ Rc) \qquad (6)$$

위의 식들에서 P_s는 셀들의 갱신주기의 합을 P_i는 셀베이스내에 저장된 각 셀의 갱신주기이다. E_t는 각 셀에 할당된 질의어 실행시간을 $Avg(E_{q_i})$는 예상 갱신비율 결정이후부터 각 셀의 질의어 실행시간 평균을 나타내며, Rc는 각 셀의 갱신횟수를 나타낸다. 각 셀들의 갱신 요청을 발생시키는 알고리즘은 (그림 2)와 같다.

```
IF current_time-last_refresh_time < cell_period
 Cell_Refresh_Count +=1
ELSE
 IF refresh_ok == true
  Cell_Refresh_Count += num_of request
 ELSE
  Cell_Refresh_Count+=num_of_request
  Last_request_time=cell_period*num_of_request
 END IF
END IF
```

(그림 2) 갱신 스케줄링 알고리즘

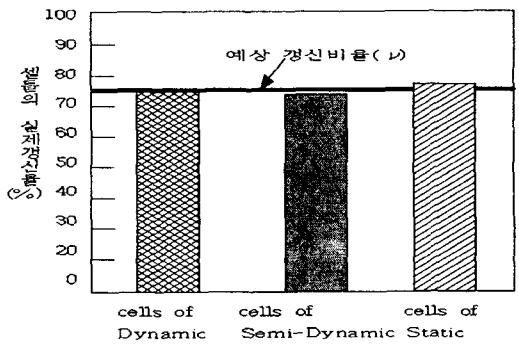
따라서 본 논문의 알고리즘은 셀의 갱신주기가 짧을수록 더 많은 실행 기회를 할당받고 갱신주기가 길수록 실행 기회를 적게 할당받게 된다. 따라서 기아상태가 특정형태의 셀에 편중되지 않으므로 위의 세 가지 형태의 웹셀들에 대하여 그룹별로 갱신비율에 대한 공정성과 적절한 신선도가 유지 될 수 있다. 갱신요청이 실행된 후 셀의 갱신비율은 증가하여 갱신된 셀의 우선순위는 결과적으로 낮아지게 된다.

3. 갱신 시스템의 구현

3.1 통합 웹사이트의 구현

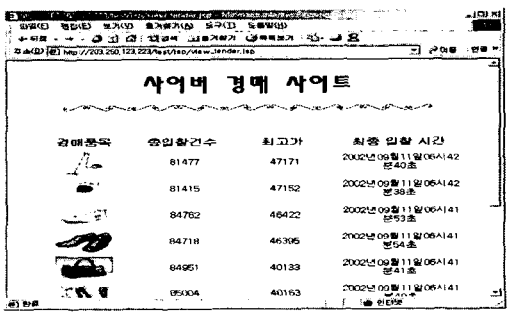
본 논문의 시스템 구현은 원격 데이터베이스 시스템(Remote Database System)으로 시스템 5대로 구축하였다. 그중 My-SQL 3.23 Database System 4대와 Oracle8 Database System 1대로 시스템을 구성하였다. 각 시스템에는 정적, 반-동적, 동적 웹사이트에 필요한 베이스데이터를 모두 구현하였다. 지역 관리 시스템(Local Management System)으로 1대의 시스템을 사용하였으며 지역관리시스템 내에는 셀베이스, 웹사이트들, Refresh Manager가 위치하게 된다. 본 논문은 정적 웹사이트로 날씨정보와 지역별 개봉영화 정보를 제공하며, 반-동적 웹사이트는 공동구매, 뉴스속보를 동적 웹사이트는 경매와 환율정보를 제공하는 웹사이트를 구현하였다. 전체 베이스 테이블 수는 100개로 구현하였다. 갱신관리자 알고리즘은 자바언어로 구현하였으며, 웹페이지 구성은 JSP로 구현하였다.

갱신 알고리즘을 구현한 결과 실제 갱신비율이 예상 갱신비율에 근접함을 (그림 3)에서 볼 수 있다.

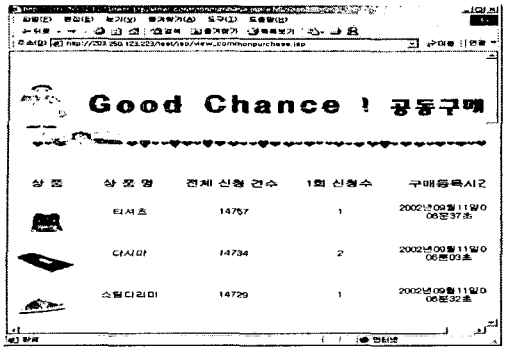


(그림 3) 그룹별 셀들의 갱신율

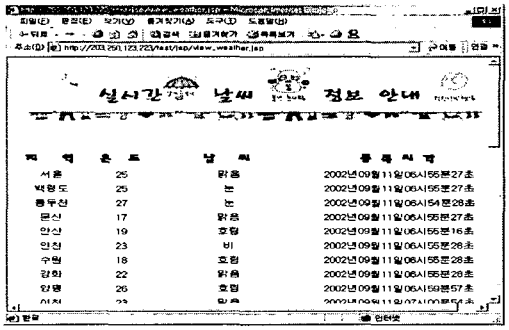
통합 구현한 웹사이트의 결과는 (그림 4), (그림 5), (그림 6)에서 각각 볼 수 있다.



(그림 4) 동적으로 구현한 사이버 경매 사이트



(그림 5) 반-동적으로 구현한 공동구매 사이트



(그림 6) 정적으로 구현한 날씨정보안내 사이트

시스템의 구현에서 동적 셀은 17개, 반동적 셀은 33개, 정적 셀은 50개로 구현한 다음 나타난 각 셀들의 시간대별 갱신 공정성은 <표 1>에서 보는 바와 같다.

<표 1> 통합시스템으로 구현한 셀들의 갱신공정성

시간(h) / 공정성 (%)	3	6	9	12	15	18	21	24
dynamic	91.0	91.9	92.3	92.0	92.1	92.2	92.3	92.2
semi -dynamic	95.2	95.3	95.4	95.3	95.2	95.2	95.3	95.3
static	98.2	97.9	97.9	98.3	98.0	98.2	98.3	98.5

3.2 기존 알고리즘과 비교 분석

기존의 MSF(Maximum Starved First)스케줄링 알고리즘[3]은 각 셀들의 기아상태의 수를 균등하게 하여 셀들의 갱신 공정성을 유지하였다. 이 알고리즘은 세가지 종류의 셀을 갱신할 경우 모든 셀의 갱신공정성은 높게 유지되지만 갱신주기가 길어질수록 셀의 신선도는 떨어진다. 반면, 본 논문의 알고리즘으로 스케줄링을 하는 경우 각 요청 셀들의 갱신 비율에 따라 갱신을 실행하기 때문에 세 가지 종류별 셀의 갱신 공정성과 모든 셀에 대한 신선도를 최적으로 유지할 수 있다.

4. 결론

본 논문에서는 현재 인터넷상에서 제공되는 정적, 반-동적, 동적 웹사이트를 통합 구현하였다. 각 웹사이트들에서 최신의 정보를 제공하기 위해 필요한 신선도 유지문제를 해결하기 위해 본 논문에서는 통합 알고리즘을 통하여 예상갱신비율을 주기적으로 재계산하는 방식을 구현하였다. 따라서 본 논문에서는 정적, 반-동적, 동적 셀들의 그룹별 갱신비율에 대한 공정성이 유지된다는 것과 세 가지 형태의 웹 셀들에 대한 적절한 신선도가 유지되는 것을 보였다.

참고 문헌(References)

[1] Haifeng Liu, Wee Keong Ng, Ee-Peng Lim, Model and Research Issues for Refreshing A very Large Website, Hong Kong, June 2000

[2] Haifeng Liu, Wee Keong Ng, Ee-Peng Lim, Keeping a Very Large Website Up-to-date: Some Feasibility Results, International Conference on Electronic Commerce and Web Technologies (EC-Web '2000, Greenwich, UK, September 2000.

[3] Haifeng Liu, Wee Keong Ng, Ee-Peng Lim, Improving the Fairness of Timely Refresh of Web Views,

[4] 황규영 외 4명 편역, 데이터베이스 시스템, 생능출판사,1997

[5] 송병호 편역, 데이터베이스 시스템, 이한출판사, 2001