김일성종합대학학보
정보과학
주체109(2020)년
제66권 제2호

웨브수집모형에 기초한 최량화된 수집일정작성방안

권세흔, 한금철

위대한 수령 **김일성**동지께서는 다음과 같이 교시하시였다.

《새로운 과학분야를 개척하며 최신과학기술의 성과를 인민경제에 널리 받아들이기 위한 연구사업을 전망성있게 하여야 합니다.》(《**김일성**전집》 제72권 292페지)

웨브수집기는 검색엔진이 사용자들에게 낡은 정보를 현시하지 않도록 국부자료기지에 보관된 자료들을 재수집하여야 한다.

선행한 방법들[1, 2]에서는 관측된 페지들의 갱신빈도수에 따라 페지들을 재수집하는 방법과 중요도에 따라 페지들을 재수집하는 방법을 제안하였다.

그러나 이 방법들에서는 국부자료기지의 무효성척도를 고려하지 못하였다.

론문에서는 국부자료기지의 무효성척도를 고려한 웨브자료수집일정작성방법을 제안하고 통합검색체계의 수집기에 적용하였다.

1. 웨브수집모형

제안한 웨브수집모형은 그림과 같다.

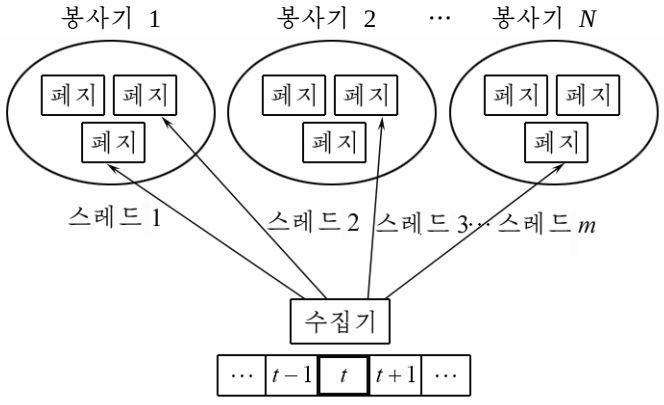

그림. 제안한 웨브수집모형

그림에서와 같이 N개의 봉사기와 m개의 수집스레드가 있다고 할 때 K는 N개의 웨브봉사기들로 이루어진 모임, W_i는 웨브봉사기 $i \in K$에 있는 웨브페지들의 모임, $W = \cup W_i$, $i \in K$는 체계에 있는 모든 페지들의 모임, p_j는 $j \in W_i$인 모든 페지의 용량이라고 정의한다.

한편 전체 재수집시간을 일정한 시간주기로 분할하고 페지내리적재는 어떤 시간주기내에서 진행하게 된다.

그리고 매 주기의 시작에서 m개의 스레드는 m개 웨브페지들을 수집하게 된다.

이때 시간주기 t의 시작시에 페지 j의 무효성척도를 $s_j(t)$로 하는데 어떤 페지가 내리적재되였을 때 그 페지의 무효성척도는 0이고 선택되지 않은 다른 페지들의 무효성척도는 주기가 끝난 다음 1만큼 증가하여 $s_j(t)+1$로 된다.

그러므로 웨브수집일정작성방법은 매 수집주기에서 페지들의 전체적인 무효성척도를 최소로 하는 봉사기 $i \in K$와 페지 $j \in W_i$를 결정하는 문제에 귀착된다.

이때 페지수집시간이 페지용량에 비례하기때문에 수집처리속도를 높이기 위해 해당한 시간주기에서 용량이 작은 페지를 선택할수 있지만 무효성척도를 감소시키기 위해 용량이 큰 페지가 내리적재되여야 하는 경우도 있다. 따라서 처리시간을 고려하면서 무효성척도를 최소로 하는 수집일정방법문제가 제기된다.

여기로부터 론문에서는 단일봉사기와 단일스레드인 경우의 수집일정작성방법을 론의하고 다중봉사기, 다중스레드로 확장하였다.

2. 단일봉사기와 단일스레드방식에서의 최량화된 수집일정작성방안

1대의 웨브봉사기에서 봉사하는 웨브페지들의 모임을 W라고 하고 $T+1$개의 수집주기들이 있다고 할 때 $g(t)$를 어떤 시간주기 t에 따르는 봉사기의 평균응답시간이라고 하면 단일봉사기에 단일스레드인 경우를 고찰하므로 어떤 시간주기 t에서는 1개의 스레드가 1개 웨브페지만을 내리적재하게 되며 따라서 시간주기 t에서의 일정작성방법을 $x(t)=(x_j(t))$, $j \in W$로 정의할수 있다. 여기서 $x_j(t)$는 주기 t에서 페지 j가 내리적재되면 1 아니면 0으로 되며 따라서 다음의 식이 성립하게 된다.

$$\sum_{j \in W} x_j(t) = 1, \ \forall t = 0, \cdots, T-1$$

한편 시간주기 $t+1$에서의 페지 j의 무효성척도를 이전 시간주기로부터 계산해보면 다음과 같은 식이 성립한다.

$$s_j(t+1) = \begin{cases} 0, & x_j(t) = 1 \\ s_j(t)+1, & x_j(t) = 0 \end{cases} \tag{1}$$

이 식을 주기에 대하여 전개하면 $s_j(t+1) = (s_j(t)+1)(1-x_j(t))$로 된다. 따라서 문제는 전반적인 페지무효성척도를 최소화하는 방안 $x^* = (x(t))$, $t = 0, \cdots, T-1$을 얻는것으로 된다.

$$\min_{x^*} \sum_{t=0}^{T} \sum_{j \in W} s_j(t) \tag{2}$$

매 주기 t의 시작점에서 전체 무효성척도를 추정해야 하므로 마지막일정작성은 $T-1$주기의 시작에서 작성되고 페지내리적재는 $T-1$주기의 마감점에서 끝나며 주기 T의 시작점에서 전체 무효성척도계산이 진행되는것으로 보아야 한다. 전체 페지의 무효성척도를 식 (1)로부터 고찰하면 다음과 같다.

$$\sum_{j \in W} s_j(t+1) = \sum_{j \in W} s_j(t) + (|W|-1) - s_{j^*(t)}(t) \tag{3}$$

여기서 $j^*(t)$는 주기 $t+1$에서 0으로 된 즉 t주기에서 내리적재된 페지를 나타낸다.

식 (3)을 $t = T-1$을 리용하여 변경하면 다음과 같은 식이 얻어진다.

$$\sum_{j\in W} s_j(T) = \sum_{j\in W} s_j(0) + T(|W|-1) - \sum_{t'=0}^{T-1} s_{j^*(t')}(t')$$

이로부터 임의의 주기 $t=0, \cdots, T-1$에 대하여 다음의 식이 성립한다.

$$\sum_{j\in W} s_j(t) = \sum_{j\in W} s_j(0) + t(|W|-1) - \sum_{t'=0}^{t-1} s_{j^*(t')}(t') \tag{4}$$

결정변수 $x(t)$를 리용하면 식 (4)로부터 다음의 식이 얻어진다.

$$\sum_{j\in W} s_j(t) = \sum_{j\in W} s_j(0) + t(|W|-1) - \sum_{t'=0}^{t-1}\sum_{j\in W} x_j(t')s_j(t') \tag{5}$$

식 (5)로부터 전체 무효성척도 $\sum_{t=0}^{T}\sum_{j\in W} s_j(t)$를 구하면 다음과 같다.

$$(T+1)\sum_{j\in W} s_j(0) + \frac{T(T+1)}{2}(|W|-1) - \sum_{t=0}^{T}\sum_{t'=0}^{t-1}\sum_{j\in W} x_j(t')s_j(t') \tag{6}$$

식 (6)이 최소로 되자면

$$\sum_{t=0}^{T}\sum_{t'=0}^{t-1}\sum_{j\in W} x_j(t')s_j(t')$$

가 최대값을 가져야 한다.

그런데 처리시간을 최소로 해야 한다는 점을 고려하면 다음의 식이 성립한다.

$$\max_{x^*} \sum_{t=0}^{T}\sum_{t'=0}^{t-1}\sum_{j\in W} x_j(t')s_j(t') - \sum_{t=0}^{T-1}\sum_{j\in W} x_j(t)p_j g(t) \tag{7}$$

식 (7)의 구조적인 의미를 음미해보자.

$T=2$일 때 식 (7)을 전개하면

$$\max_{x^*} \sum_{j\in W}(2s_j(0) - p_j g(0))x_j(0) + \sum_{j\in W}(s_j(1) - p_j g(1))x_j(1) \tag{8}$$

로 되고 $T>2$일 때에는

$$\begin{aligned}\max_{x^*} &\sum_{j\in W}(Ts_j(0) - p_j g(0))x_j(0) + \sum_{j\in W}((T-1)s_j(1) - p_j g(1))x_j(1) + \cdots + \\ &+ \sum_{j\in W}(s_j(T-1) - p_j g(T-1))x_j(T-1)\end{aligned} \tag{9}$$

로 된다.

식 (9)로부터 수집일정작성방법은 어떠한 시간주기 t에서 재수집해야 할 페지를 선택하는 문제에 귀착되며 시간주기 t에서의 페지선택은 다른 시간주기들과는 독립이라는 것을 알수 있다. 따라서 수집일정작성방법에서는 시간주기 t에서의 페지선택 $x(t)$를 다음의 식이 최대로 되도록 결정하면 된다.

$$\sum_{j\in W}((T-t)s_j(t) - p_j g(t))x_j(t)$$

우의 식과 $\sum_{j\in W} x_j(t)=1$로부터 다음과 같은 식이 성립한다.

$$j^*(t) = \arg\max_{j \in W}((T-t)s_j(t) - p_j g(t))$$

이 식은 매 시간주기 t에서 봉사기의 평균응답시간 $g(t)$가 주어질 때 재수집해야 할 페지선택은 $s_j(t)$와 p_j값에 의하여 결정된다는것을 보여준다.

결과적으로 수집일정작성방법에서는 모든 페지들의 용량이 같은 특수한 경우($\forall j \in W,\ p_j = p$) 수집기는 무효성척도 $s_j(t)$가 최대인 페지 j를 선택하며 모든 페지들이 같은 무효성척도를 가지는 경우에는 용량이 최소인 페지 j를 선택한다.

3. 모 형 확 장

① 다중봉사기, 단일스레드일 때의 모형확장

$N>1$개의 웨브봉사기들의 모임 K를 고려할 때 매 봉사기들은 평균응답시간 $g_j(t)$값을 가지게 되며 매 시간주기 t에서 오직 1개 스레드만($m=1$)이 웨브봉사기로부터 페지를 내리적재하게 되므로 수집일정작성방법은 모든 봉사기들의 전체 무효성척도가 최소인 방안 $x^* = x(t),\ \ t=0,\ \cdots,\ T-1$을 구하는 문제로 된다.

$$\min_{x^*} \sum_{t=0}^{T} \sum_{i=1}^{N} \sum_{j \in W_i} s_j(t) \tag{10}$$

매 시간주기 t에서 어떤 봉사기의 웨브페지를 내리적재하는 스레드는 1개이므로 다음의 식이 성립한다.

$$\sum_{i \in K} \sum_{j \in W_i} x_j(t) = 1,\ \forall t = 0,\ \cdots,\ T-1 \tag{11}$$

우의 방법과 류사한 방법으로 식 (10)을 전개하면 식 (12)가 성립한다.

$$\max_{x^*} \sum_{t=0}^{T} \sum_{t'=0}^{t-1} \sum_{i=1}^{N} \sum_{j \in W_i} x_j(t') s_j(t') \tag{12}$$

앞에서 언급한 방법과 류사하게 다중웨브봉사기인 경우의 수집일정작성방법도 페지재수집시간을 최소로 하는 측면을 고려해야 한다.

따라서 다중봉사기를 위한 수집일정작성방법은 다음과 같다.

우선 봉사기들의 모든 페지들의 무효성척도가 같은($\forall i \in K,\ \forall j \in W_i, s_j(t) = s$) 경우 수집기는 $p_j g_i(t)$가 최소인 봉사기 i와 페지 j를 선택하며 모든 페지들의 용량이 같은 경우($\forall i \in K,\ \forall j \in W_i,\ p_j = p$)에는 $g_j(t)$가 최소인 봉사기에서 페지를 우연적으로 선택한다.

다음으로 봉사기들의 모든 페지들의 용량이 같은($\forall i \in K,\ \forall j \in W_i,\ p_j = p$) 경우에는 $s_j(t) - g_i(t)$가 최대인 봉사기 i와 페지 j를 선택하며 t주기에서 $g_j(t) = g(t)$이면 $s_j(t)$가 최소인 페지를 선택한다.

② 다중봉사기, 다중스레드일 때의 모형확장

수집기가 $m>1$개의 스레드를 리용하므로 매 시간주기 t에서 m개의 페지들을 재수집해야 한다. 이 경우 식 (11)로 다음의 식이 성립하게 된다.

$$\sum_{i \in K} \sum_{j \in W_i} x_j(t) = m,\ \forall t = 0,\ \cdots,\ T-1 \tag{13}$$

따라서 다중봉사기와 다중스레드인 경우의 수집일정작성방법에서는 매 시간주기 t의 시작점에서 다음의 식이 최대인 m개의 봉사기－페지쌍을 선택한다.

$$(T-t)s_j(t)-p_jg_i(t)$$

4. 실험 및 성능평가

제안한 방법을 통합검색체계에서 수집해야 할 홈페지 10개를 대상으로 하여 적용하였다.

실험을 진행한 후 웨브페지수집자료기지의 페지평균무효성척도는 표와 같다.

표. 웨브페지수집자료기지의 페지평균무효성척도

	제안한 방법	선행방법[1]	선행방법[2]
페지평균무효성척도/min	235.4	526.5	897.7

표로부터 제안한 방법이 웨브페지수집자료기지의 페지평균무효성척도를 선행방법[1]보다는 2.24배, 선행방법[2]보다는 3.81배 제고하였다는것을 알수 있다.

맺 는 말

이미 수집된 페지들의 무효성척도를 고려한 최량화된 재수집일정작성방안을 제안하고 실험을 통하여 통합검색체계의 수집기의 성능을 검증하였다.

참 고 문 헌

[1] Q. Tan et al.; ACM Trans. Inf. Syst., 28, 4, 1, 2010.

[2] S. U. Khan et al.; Cluster Comput., 16, 1, 65, 2013.

주체109(2020)년 2월 5일 원고접수

Optimal Web Page Download Scheduling Policy Based on the Web Crawling Model

Kwon Se Hun, Han Kum Chol

We proposed an optimal scheduling method to minimize the total staleness of pages in the repository of a web crawler and proved its efficiency through the experiments with the integrated search system.

Keywords: crawling, staleness, scheduling policy