효율적 웹 기반 VOD 서비스를 위한 에이전트 시스템[1)]

이경희* · 한정혜** · 김동호**
충북대학교 전자계산학과* · 청주교육대학교 컴퓨터교육학과**

요 약

VOD 서비스를 제공하는 서버는 클라이언트에게 좋은 서비스를 제공하기 위하여 요청을 처리할 수 있는 여러 개의 부본서버를 이용한다. 또한 시공간적으로 다양한 환경으로부터 요청된 클라이언트의 요구를 가장 효과적으로 처리할 수 있는 서버를 선택하기 위한 다양한 서버 선택 알고리즘들이 제안되었다. 선택 알고리즘은 각 부본서버의 성능을 평가할 수 있는 척도에 의하여 클라이언트의 요청을 적절하게 분산시켜야 서버 부하의 균형이 이루어지고 서비스의 질이 향상된다. 이때 척도를 어떻게 결정하느냐 하는 것이 선택알고리즘의 성능을 결정짓는 중요한 요소이다. 본 논문에서는 서버의 성능에 따라 서비스 요청을 분산시킴으로써 QoS를 향상시키는 데 중점을 둔 에이전트 시스템을 제안하고자 한다. 이 시스템은 부본서버 성능 분석자, 지식베이스, PCI 알고리즘기반 오토마타로 구성하였다. 학습자들의 요청에 응답한 각각의 서버로부터 수집한 로그를 분석하고, 조사하여 지식베이스를 생성하고 이로부터 추출한 규칙에 의하여 서비스를 제공할 서버를 선택한다. 여러 개의 부본서버 중 사용자 요청에 대하여 가장 좋은 서비스를 제공할 것으로 기대되는 서버를 동적으로 선택한다.

An Agent System for Efficient VOD Services on Web

Kyung-Hee Lee* · Jeong-Hye Han** · Dong-Ho Kim**

ABSTRACT

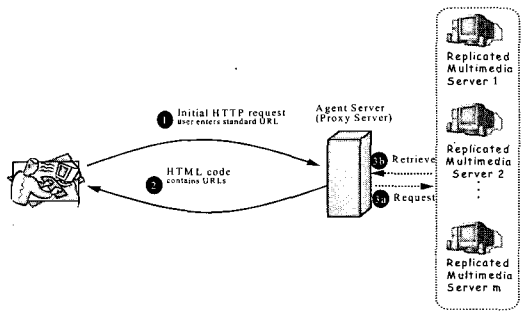
Most of the existing algorithms try to disseminate the multimedia contents of internet service provider(ISP), without taking into account the needs and characteristics of specific websites including e-Learning systems with web-based educational contents. Sometimes the client must select the best one among the replicated repositories. However, this is a less reliable approach because clients' selections are made without prior information on server load capacity. In this paper we propose an agent system inspired by the need of improving QoS of delivering web-based educational multimedia contents without incurring long access delays. This agent system consists of three components, Analyzer, Knowledge Base, and Automaton embedded the capacity algorithm. It analyzes and investigates traffic information collected from individual replicated server by learners' requests, and selects a server which is available and is expected to provide the fastest latency time and the lowest loaded capacity, and achieves high performance by dynamic replicating web resources among multiple repositories..

Ⅰ. 서 론

초고속 인터넷과 고성능 PC 보급으로 웹 서비스 사용자의 폭이 확대되고 다양한 분야에서 웹을 이용한 서비스를 제공하고 사용하게 되었다. 많은 전문가들이 전세계 사람들과 협력하기 위한 방편으로 웹을 사용하고 있다. 특히 정보전달, 교육, 홍보, 문화교류 등에 효과적인 멀티미디어 데이터를 웹을 통하여 전송하는 기술이 각광받고 있다. 멀티미디어 데이터는 어학, 영화, 음악 등으로 웹을 통하여 직접 동영상 등을 볼 수 있는 주문형 영상정보 시스템으로 현재 초기 단계인 VOD (Video on Demand)의 형태로 대부분 서비스되고 있으며, 이러한 VOD의 가장 큰 장애물은 비디오 전송이 요구되는 대량의 데이터를 신속하게 처리할 만한 네트워크 기반 시설이 아직은 부족하며, 멀티미디어 데이터 서비스의 상업화에 따라 QoS(Quality of Service)가 더욱 중요해졌다. 특히 멀티미디어 VOD의 경우는 데이터가 연속적으로 재생되어야 하기 때문에 클라이언트가 수동으로 서비스를 요청할 서버를 선택하는 것은 바람직하지 못하다. 좋은 서비스를 제공할 서버를 선택할 수 있는 능력이 사용자에게 있거나 혹은 서버간에 트래픽을 조정할 중재자가 있어야 한다.

따라서 이러한 멀티미디어 데이터 서비스의 QoS를 향상하기 위하여 이와 관련된 여러 연구들이 진행되고 있는데, 크게 분류하여 서버에 여러 개의 버퍼를 두어 데이터 스트림을 버퍼링하는 방법[3, 5], 아래 (그림 1)과 같이 하나의 서버에서 모든 클라이언트의 요청을 처리하는 것이 아니라 클라이언트를 가까운 부본서버(Duplicated Server)로 분산시켜 서비스를 제공하도록 하는 방법 등이 있다. 이미 여러 웹 사이트에서 성능향상과 신뢰도 향상을 위하여 웹 문서를 분산시키고, 부본서버를 통해 서비스하는 방법을 사용하고 있다[6].

그러므로 본 연구에서는 서비스 지연시간을 최소화함으로써 VOD서비스의 QoS 향상시킬 수 있는 에이전트 시스템을 제안한다. 이 시스템은 클라이언트의 요청을 접수한 후 가장 좋은 서비스를 제공할 수 있을 것으로 기대되는 부본서버를 선택하여 요청을 전달하는 기능을 한다. 에이전트 시스템은 각 부본서버의 성능을 분석하는 부분, 서버의 정보를 저장하는 부분 그리고 서버를 선택하는 부분으로 구성된다. 클라이언트의 요청에 능동적으로 서비스할 수 있기 위한 전제조건은 웹로그를 분석한 트래픽 사전정보를 통해 각 부본서버의 부하량을 체크하고, 이후에 발생하는 클라이언트의 요청을 분산시킬 수 있는 능력을 갖는 것이다.



(그림 1) 부본서버를 이용한 일반적인 VOD서비스 아키텍처

본 논문의 구성은 2절에서 부본서버를 이용한 트래픽 신뢰도 향상에 대한 관련 선행연구와 알고리즘을 요약하고, 3절에서는VOD서버의 용량을 반영할 수 있는 메트릭으로서 공정능력지수(Process

1) 이 논문은 2001년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2001-003 -E00261)

Capability Index:PCI) 알고리즘을 이용한 동적 부본서버 동적선택 에이전트를 시스템을 제안한다. 마지막 4절에는 결론과 향후 연구를 제시하겠다.

Ⅱ. 관련연구

클라이언트의 요청에 가장 적합한 부본서버 하나를 선택하기 위한 기준으로 근접척도(approximity metric)라는 것이 있는데, 이러한 근접척도는 지리적 위치, 네트워크 연결상태, 트래픽 부하 등 다양하게 설정될 수 있다. 먼저 [2]은 지리적 거리, 연결 홉(hop)의 개수, 난수발생, 핑 요청의 라운드 트립 시간(round trip measurement)의 평균값을 제안하였다. 그러나 [6]은 홉의 수나 핑 요청의 라운드 트립 시간이 웹 문서의 크기가 아주 작을 경우에는 적당하지만, 근접한 부본서버를 결정하는 일반적으로 좋은 척도가 아님을 보였다. 즉, 요청 후 완전한 문서가 오기까지의 HTTP 요청 응답시간(response time)보다 요청 후 첫 바이트가 오기까지의 HTTP 요청·반응시간(latency time)의 사용이 더 효과적임을 제안했으며, 〈표1〉과 같이 두 척도간의 상관계수값 사례결과도 보였다. 직관적으로 HTTP 반응시간과 HTTP 응답시간의 상관관계는 매우 높을 것이고, HTTP 반응시간과 문서크기의 상관관계는 거의 없을 것이다.

〈표 1〉 근접척도간의 상관관계

알고리즘 척도	HTTP 응답시간	HTTP 반응시간
핑	0.51	0.76
홉의 개수	0.16	-
HTTP 응답시간	-	0.73

HTTP 반응시간을 이용한 부본서버 동적선택 알고리즘을 정리하면 다음과 같다[6][7].

- 병렬(parallel):모든 부본서버에 매번 요청을 보내 가장 빠른 HTTP 반응시간을 보이는 부본서버 선택. 가장 효과적이나 동시 사용자에게 같은 부본서버를 선택하게할 위험.

  확률(probabilistic):가장 빠른 HTTP 반응시간을 갖는 부본서버가 선택될 확률을 크게 배정. 부본서버의 정보를 갱신할 추가요청이 필요없으나, 최선이 아닌 부본서버에 보내지는 요청회수를 통제할 수 없음.

  $\hat{P}(S_i)=\frac{k}{\bar{l}_i}$, 단, $\bar{l}_i$는 평균반응시간,

  k는 $\sum_{i=1}^{m} P(S_i)=$을 만족하는 상수

  재시도(refresh):제한된 시간에 HTTP 반응시간내 반응이 오면 그 부본서버를 선택하고, 초과하면 새로운 HTTP 반응시간 값을 얻는다. 추가 요청을 생성하는 단점이 있으며, 리플래시 요청 회수는 HTTP 반응시간값으로 조정할 수 있음.
- 확장된 재시도(extended refresh):여러 부본서버의 HTTP 반응시간 평균은 유사하나 서비스 질을 의미하는 분산은 크게 다르므로, S-percentile기법과 EWMA, EWMV 모형을 이용하여 HTTP 반응시간값을 추정하는 재시도 알고리즘의 확장[7].

위의 동적선택 알고리즘 외에도 동일확률배분의 고정 알고리즘과 홉의 개수를 이용한 정적 알고리즘은 [2]를 참고하면 된다.

이러한 정적·동적선택 알고리즘들은 기존의 서버운영 로그결과 분석이 선행되어야 하므로, 서비스를 처음 시작하는 경우에는 적용하기 적당치 않

은 것도 있다. 이런 경우 문서의 인기도인 이용접속 확률분포에 의하여 부본서버를 선택하거나[1], 문서의 멀티미디어 개체 크기와 문서의 이용접속 확률행렬을 고려하여 부본에서 다운로드 시간을 최소화하기 위한 비용모형(cost model)을 제시하였다[8]. 즉, 멀티미디어 컨텐츠의 크기가 크거나 스트리밍 데이터인 경우는 HTTP 반응시간 평균과 같은 근접척도를 이용한 이러한 알고리즘 보다는 서비스될 해당 부본서버의 하드웨어 처리성능에 따른 편차가 고려된다면 더욱 신뢰도가 높아질 것이다. 따라서 다음 절에서는 크기가 큰 VOD용 멀티미디어 서버의 QoS를 높이기 위하여, 하드웨어 성능 및 웹로그 자료에 근거한 공정능력지수를 근접척도로 하는 PCI 알고리즘을 제안하겠다.

Ⅲ. 동적부본서버 선택 에이전트 시스템

3.1 에이전트 시스템

본 절에서는 부본서버를 이용한 VOD 서비스의 QoS를 높이기 위한 에이전트 시스템 구성과 각 콤포넌트의 알고리즘을 제안하고자 한다.

먼저 m개의 부본서버 $\vec{R}=(R_1, R_2, \cdots, R_m)$가 있다고 했을 때, 임의의 지리적 위치에서 VOD 서비스를 요청한 클라이언트 C에 대하여 최적의 서비스를 위해서는 다음 〈식 1〉의 Q를 최소화 하는 부본서버를 선택하도록 제공하는 것이다. 즉, 클라이언트와 서버간의 네트워크 부하량이 최적인 서버와 서버부하가 최소인 부본서버를 선택하는 것이다.

$$Q=\min(Network\ Load)+\min(RL_1, RL_2, \cdots, RL_m) \quad \langle 식1 \rangle$$

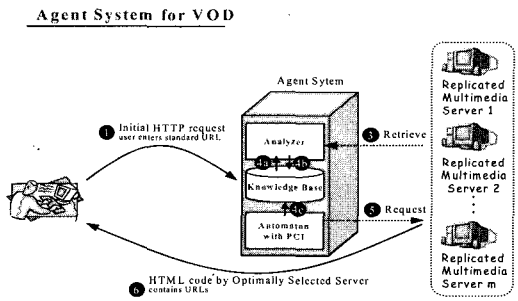
단, RL_i : i번째 부본서버 부하량, $i=1,2,\cdots,m$

이때 네트워크 부하는 [2]와 같이 접속하는 클라이언트의 지리적 위치 등에 의해 측정될 수 있으나 외부요인으로 서버관리자가 제어할 수 없으므로, 이 항을 상수로 가정한다면 각 부본서버 중 부하량이 최소인 서버를 선택하는 것이 최적의 선택이 될 수 있다.

[1]은 〈식1〉의 두 번째 항을 문서의 요청확률, 단위시간당 서비스된 문서의 크기(byte), 서버저장능력의 함수로 보고 최소화를 위한 알고리즘, [8]은 서비스하는 문서에 포함된 멀티미디어 오브젝트의 크기에 따른 전송비용을 이용한 알고리즘, [6]은 RLi를 HTTP 반응시간의 함수로 보고 평균을 이용한 알고리즘을 제안하였다.

그러나 이러한 알고리즘은 부본서버들간의 하드웨어적인 성능을 수학적으로 고려하지 않은 것이며, 부본서버들간의 하드웨어적 성능차가 매우 커서 클라이언트 요청에 의한 HTTP 반응시간 평균값의 변화가 매우 심한 경우에는 효과적이지 않다. 따라서 본 연구에서는 RLi를 하드웨어적 성능과 HTTP 반응시간의 변화를 모두 고려한 PCI 알고리즘을 포함한 동적부본서버 선택 에이전트 시스템을 제안하고자 한다.

다음 (그림 2)는 VOD 부본서버의 성능을 평가할 수 있는 척도를 이용한 서버의 PCI값에 의해 부본서버를 선택할 수 있는 에이전트 시스템의 구조이다.



(그림 2) 제안된 VOD서비스 에이전트 시스템 아키텍처

이 시스템은 클라이언트의 요청을 접수한 후 가장 좋은 서비스를 제공할 수 있을 것으로 기대되는 부본서버를 선택하여 요청을 전달하는 기능을 한다. 에이전트 시스템은 각 부본서버의 성능과 상태를 분석하는 부분(Analyzer), 서버의 정보를 저장하는 부분(Knowledge Base) 그리고 서버의 성능에 따라 자동으로 서비스할 서버를 선택하는 부분(Automaton with Capacity Algorithm)으로 구성된다.

Analyzer는 m개의 부본서버로부터 주기적으로 웹로그를 수집하여 서버성능을 분석한다. 분석된 결과는 지식베이스에 저장하고 PCI알고리즘은 지식베이스에 저장된 각 부본서버의 정보로부터 근접척도를 계산하여 현재 서비스를 요청한 클라이언트에게 가장 좋은 서비스를 제공할 수 있는 부본서버를 선택한다.

3.2 PCI 알고리즘

VOD 컨텐츠들을 똑같이 갖고 있는 각 부본서버들에게 특정 문서의 요청이 이루어졌을 때, 기존의 알고리즘에 의해서 해당 부본서버가 선택될 것이다. 그러나 부본서버들간의 하드웨어적 성능의 차가 큰 경우 각 부본서버들간의 서비스 처리 속도는 매우 달라 단순히 HTTP 응답시간 평균값만을 고려한다면 효과적이지 못하다. 실제로 많은 상업적 사이트의 부본서버들은 예산에 따라 추가 구입 및 업그레이드가 되므로 하드웨어 성능차가 큰 경우가 많다.

따라서 클라이언트가 요청한 임의의 고정된 시점에 있어서, 서비스를 수행중인 부본서버 RLi는 다음과 같은 함수로 고려할 수 있으며, 이 함수를 최소인 부본서버를 선택하면 되는 것이다.

$$RL_i = f(HW, SLM, SLS) \qquad \langle 식2 \rangle$$

단, HW 하드웨어성능,
SLM : 응답시간평균, SLS : 응답시간편차

이때 이러한 함수를 반영하는 메트릭이 요구되는 것인데, HTTP 응답시간 변인에 대하여 다음 〈식 3〉의 PCI를 적용한다면 〈식 2〉의 세 입력변수를 대표할 수 있다.

$$C_p = \frac{USL - LSL}{6\sigma} \qquad \langle 식3 \rangle$$

단, $USL(LSL)$: 최대(소)허용 $HTTP$ 응답시간

〈식 3〉의 추정량은 스틸링의 공식에 의하여 n (≥30)이 충분히 클 때 일반적인 불편추정량을 사용할 수 있으며[4], 이 추정량에 대한 관측값이 (그림 2)의 지식베이스에 저장되며 Automaton에 의해 계산되어 최종 클라이언트에게 최적의 서버를 선택하여 서비스를 하는 것이다.

이 C_p의 추정량값이 크면 해당 부본서버의 HTTP 응답시간의 변동이 작으므로 해당 부본이 선택될 확률을 크게 하고, 반대로 추정된 C_p값이 작으면 해당 부본서버의 HTTP 응답시간의 변동이 크므로 그만큼 하드웨어적 성능과 접속량 처리가 늦으므로 QoS를 좋게 하기 위하여 해당 부본서버가 선택될 확률을 작게 하는 것이다.

PCI알고리즘 기반 오토마타에 포함되는 알고리즘은 다음 (그림 3)과 같다.

공정능력지수 C_p 추정량의 분포는 정규분포 가정하에서 조차 매우 복잡하므로, 지수분포와 같은 분포에 관계없는 추정량으로 붓스트랩 알고리즘을 적용할 수도 있을 것이다[4]. 또한 [4]에서의 다양한 붓트스랩 퍼센타일 기법을 이용하여 지수분포의 모수를 추정할 수도 있다.

```
FUNCTION Calcul_PCI()
{
    int i;
    CONST float h;
    For (i=1; in; i++){
        C_pi = h/6 S_i;
        If ( C_pi >1) then P( S_i )=k/ l_i
        Else P( S_i )=k*( C_pi / l_i );
    } /* k is calculated under the constraint
        the sum of probabilities equals to 1*/
} /* End of FUNCTION Calcul_PCI */
FUNCTION Server_Selection()
{
  int i;
  /* # of replicated repositories  */
  i = Probabilistic_Selection();
  if ( C_i > M_i ) {i=Probabilistic_Selection();}
  else return i;
}
```

(그림 3) PCI C_p 알고리즘 Pseudo 코드

Ⅳ. 결 론

부본서버를 이용하여 멀티미디어 컨텐츠 서비스를 제공하는 경우에 사용하는 다양한 선택알고리즘을 살펴보았다. 멀티미디어 문서의 이용분포를 살펴보면 크기가 작은 경우가 빈번히 사용되는 지수분포를 갖는 것에 의해, 작은 문서의 경우는 HTTP 반응시간을 이용한 부본서버를 동적으로 선택하는 알고리즘이 제안되었다.

그러나 VOD 서비스를 제공하는 많은 사이트들의 부본서버들은 하드웨어의 구입시기와 예산 때문에 하드웨어적 성능의 차가 크며, 이 성능차를 고려한 추가적인 메트릭이 요구된다.

따라서 본 논문에서는 VOD와 같은 멀티미디어 컨텐츠에 대한 부본서버의 하드웨어 성능에 따라 HTTP 반응시간이 매우 다른 경우에 유용한 부본서버 동적선택 알고리즘을 제안하였다.즉, HTTP 반응시간의 평균뿐만 아니라 편차를 고려하여 QoS를 높일수 있는 공정능력지수 PCI C_p를 이용한 알고리즘을 제안하였다. VOD 서비스 처리능력이 떨어지는 부본서버에는 작은 선택확률을 부여하고, VOD 서비스 처리능력이 안정적인 부본서버에는 큰 확률을 부여하는 것이다.

향후 연구로는 본 에이전트 시스템의 Analyzer에 사용될 추정량이 실시간(real time)으로 계산될 경우에 편의(bias)를 줄이기 위한 붓스트랩(bootstrapping) 기법을 이용함으로써 그 응용범위를 확대하고자 한다.

참고문헌

[1] Azer Bestavros, "Demand-based Document Dissemination to Reduce Traffic and Balance Load", Proceedings of SPDP, 1995.

[2] M.E. Crovella, R.L. Carter, "Dynamic server Selection in the Internet", proceedings of the Third IEEE workshop on the Architecture and Implementation of High Performance Communication Subsystems, 1995.

[3] Harvey M. Deital, "An Introduction to Operating Systems", Addison Wesley, 1983.

[4] Han J.H., Cho J.J, Leem C.S, "Bootstrap

Confidence Limits for Wright's Cs", Communications in Statistics: Theory and Methods, Vol. 29, No. 3, pp. 485-505, 2000.

[5] David A. Patterson and John L. Hennessy, "Computer Architecture a Quantitative Approach", Morgan Haufmann Publisher, 1996.

[6] Mehmet Sayal, Yuri Breitbart, Scheu -ermann, "Radek Vingralek, Selection Algorithms for Replicated Web servers", Proceedings of the Workshop on Internet Server Performance, 1998.

[7] Radek Vingralek, Yuri Breitbart, Mehmet Sayal, Peter Scheuermann., "Web++:A System for Fast and Reliable Web Service", Proceedings of the USENIX Annual Technical Conference, 1999.

[8] Thanasis Loukopoulos and Ishfaq Ahmad, "Replicating the Contents of a WWW Multimedia Repository to Minimize Download Time", Proceedings of the 14th International Parallel and Distributed Processing Symposium, 2000.

이 경 희

1999년 충북대학교 전자계산학과 (이학석사)

1999년 - 현재 충북대학교 전자계산학과 박사과정

2000년 - 현재 (주)엔슬래시닷컴 선임연구원

2001년 - 현재 극동정보대학 전산정보처리과 겸임교수

관심분야 : 전자상거래, 모바일 서비스, XML

한 정 혜

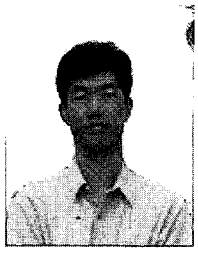
1998년 충북대학교 전자계산학과 (이학박사)

1998년-1999년 연세대학교 산업시스템공학과 포닥연구원

1999년-2001년 행정자치부 국가전문행정연수원 통계연수부 전산교육 전임교수

2001년 - 현재 청주교육대학교 컴퓨터교육과 교수

관심분야는 멀티미디어통신, EC, 데이터마이닝



김 동 호

1986년 서울대학교 계산통계학과(이학사)

1988년 서울대학교 계산통계학과(계산학 석사)

1999년 서울대학교 전산과학(이학박사)

1990년 - 현재 청주교육대학교 컴퓨터교육학 교수

관심분야 : 자연언어처리, 기계번역시스템, 소프트웨어공학, 객체지향시스템, 웹기반교육, 정보기술과 교사교육 등