웹 서버에서의 품질 정보를 이용한 웹 서비스 스케줄링 기법

김동준° 이상규 한상용
중앙대학교 컴퓨터공학과
{djkim°, sklee}@archi.cse.cau.ac.kr, hansy@cau.ac.kr

A Web Services Scheduling Strategy using Quality Information on Web Server

DongJoon Kim°, SangKyu Lee, SangYong Han
Dept. of Computer Science and Engineering, Chung-Ang University

요 약

최근에 많은 기업들은 XML 메시지 기반의 분산 환경에서 이기종 간의 표준으로 자리잡고 있는 웹 서비스를 도입하여 B2C 뿐만 아니라 B2B에 거쳐서 기업과 기업간 동적인 비즈니스를 수행하고 있다. 하지만, 웹 서비스의 도입이 성공적으로 이루어지기 위해서는 차별화된 서비스 품질을 보장해 주어야 한다. 현재 대부분의 웹 서비스에서는 메시지에 대하여 차별화를 두고 있지 않으며, 기존의 웹 서버들은 웹 서비스 제공자와 사용자간에 체결된 서비스 수준 계약에 대한 품질 요소들을 적절하게 반영하지 못하고 있다. 본 논문에서는 차별화 서비스에 사용되는 응용 수준에서의 적합한 품질 요소를 분석하고, 이러한 품질 요소를 이용하여 웹 서비스 메시지를 처리하는 프로세스에 대하여 서비스 수준 계약을 최대한으로 만족시키기 위하여 동적으로 우선순위를 할당할 수 있는 스케줄링 기법을 제안한다.

1. 서 론

인터넷은 많은 사람들이 보편적으로 사용하고 있으며, 최근에 많은 기업들은 XML 메시지 기반의 분산 환경에서 이기종 간의 표준으로 자리잡고 있는 웹 서비스를 도입하여 B2C뿐만 아니라 B2B에 거쳐서 기업과 기업간 동적인 비즈니스를 수행하고 있다. 하지만, 웹 서비스의 도입이 성공적으로 이루어지기 위하여 반드시 필요한 것은 웹 서비스 사용자가 만족할 수 있는 서비스 품질을 보장해야 한다는 것이다. 웹 서비스의 품질을 보장하기 위하여 서비스 공급자는 다양한 서비스 수준에 따라 즉, 차별화된 웹 서비스를 제공할 수 있어야 하며, 이러한 다양한 서비스 수준을 보장하기 위해서는 반드시 사용자와 제공자간의 서비스 수준 계약(SLA: Service Level Agreement)이 필요하다[1]. 서비스 수준 계약은 웹 서비스 사용자와 제공자 사이의 책임 관계를 정의하고 제공하는 서비스의 신뢰성을 보장하기 위하여 이들 상호간에 체결하는 것으로, 웹 서비스 제공자는 서비스 수준 계약에 체결된 웹 서비스 품질을 보장할 수 있어야 한다.

하지만, 현재의 웹 서비스 기술 표준 단체들은 웹 서비스 품질을 평가하기 위한 요소 혹은 평가된 정보를 기술하기 위한 목적의 언어에 대하여 표준을 만들어내지 못하고 있으며, 단지 몇몇 벤더들이나 학계에 의하여 개별적인 명세화와 연구가 진행되고 있다[2][3]. 웹 서비스 품질은 성능, 신뢰성, 가용성, 보안 등과 같이 기능적 서비스를 제공할 때의 서비스의 수준을 나타내며, 품질에 대한 분류 기준에 따라서 다양하게 존재할 수 있다.

한편, IETF(Internet Engineering Task Force)에서는 네트워크 수준에서 차별화된 서비스의 품질을 보장하기 위하여 차별화 서비스(DiffServ: Differentiated Service)를 제안 하였다[4]. 하지만 차별화 서비스에서는 인터넷을 경유한 종단간(end-to-end) 전송 품질의 보장이 되지 않으며, 통신망의 각 구간별 차별화 패킷 전송 처리 기능만이 구현되었다. 이러한 점을 보완하기 위하여 최근에는 웹 서버에서 차별화 서비스를 제공하려는 연구들이 많이 진행되고 있다[5][6][7]. 특히 웹 서비스에서는 SOAP 메시지의 파싱시간과 비즈니스 로직을 실행하는 시간이 존재하기 때문에 네트워크 지연시간보다 웹 서버에서의 지연시간이 커지는 경향이 많이 있으며, 이 때문에 웹 서버에서의 차별화 서비스는 더욱더 중요하다. 이러한 연구들은 네트워크 수준뿐만 아니라 애플리케이션 수준의 서비스를 위하여 사용자의 요청에 따라 프로세스들을 스케줄링하는 방안들을 연구하고 있다. 현재 대부분의 웹 서버들은 FIFO, static Priority 등의 스케줄링 방식을 사용하고 있다. 하지만, 이들 방식은 과거에 서비스되었던 품질 정보에 대한 성능 평가가 반영되지 못하고 있으며, 각 상황에 맞게 동적으로 우선순위를 할당하지 못하게 되어 우선순위가 낮은 프로세스에 대하여 기근(starvation) 문제를 발생 시킨다.

본 논문에서는 차별화된 웹 서비스를 애플리케이션 수준에서 제공하기 위하여 다양하게 존재하고 있는 웹 서비스 품질 요소들을 분석하여 차별화 서비스에 사용되는 적합한 웹 서비스 품질 요소를 정의하고, 이러한 품질 정보를 이용하여 각 상황에 맞게 메시지를 처리하는 프로세스에 대하여 동적으로 우선순위를 할당함으로써 서비스 수준 계약을 최대한 만족할 수 있는 스케줄링 기법을 제시한다.

2. 웹 서비스 품질 요소

웹 서비스의 중요성이 증대됨에 따라 서비스의 품질은 서비스의 사용 빈도와 상호간의 신뢰성에 중요한 영향을

이치게되며, 이는 서비스 제공자의 성공을 위한 중요한 요소가 되어가고 있다. 현재의 웹 서비스 기술 표준화 단체들은 웹 서비스 품질을 평가하기 위한 요소 혹은 평가된 정보를 기술하기 위한 목적의 언어에 대한 노력이 미약한 실정이며 단지 몇몇 벤더들에 의하여 개별적인 명세화가 진행되고 있다. 웹 서비스 품질은 성능, 신뢰성, 가용성, 보안 등과 같이 기능적 서비스를 제공할 때의 서비스의 수준을 나타내며, 품질에 대한 분류 기준에 따라서 다양하게 존재할 수 있다[2][3]. 하지만, 우선순위를 결정하기 위하여 너무 많은 품질 요소를 반영하기에는 정확성 및 성능적인 측면에서 무리가 따른다. 이에 본 논문에서는 서비스 수준 계약에서 가장 많이 사용되며 모니터링하기가 용이한 성능 및 안전성 측면의 품질 요소를 사용하며, 이 품질 요소는 사용자의 기대 성능을 잘 반영하고 있다. 다음은 고려되는 성능 및 안정성 측면의 품질 요소에 대하여 설명한다.

○ 성능 측면의 품질 요소

성능 측면의 품질 요소를 반영함으로써 우선순위를 부여함으로써 높은 성능을 요하는 서비스 사용자에게 차별화된 서비스를 제공할 수 있다.

- 응답 시간(Response Time) : 다양한 타입의 요청에 대해서 서비스가 응답하는데 소요되는 시간을 의미한다. 응답 시간은 단위 시간에 대한 응답 비율 혹은 동시 요청수의 관점에서 측정할 수 있다.
- 처리량(Throughput) : 주어진 시간 동안 제공된 서비스 요청 수를 의미하며, 서비스가 처리할 수 있는 요청 비율로 나타내거나 품질 정보 측정 시 최대 처리량 혹은 요청 부하에 대한 처리량 변화 함수를 통해 표현한다.

○ 안정성 측면의 품질 요소

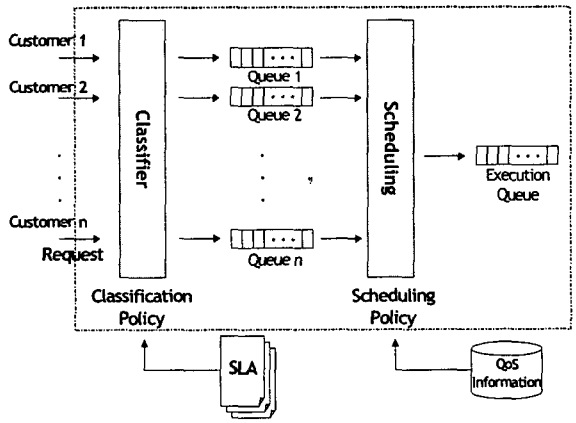
안정성 측면의 품질 요소는 웹 서비스의 사용 가능성에 관한 품질 요소이며, 이 역시 차별화 서비스가 이루어져야 한다. 이 품질 요소는 웹 서버에서 우선순위를 할당하는 것과 관계 없어 보일 수 있으나, 사용자 관점에서는 웹 서비스가 사용 가능한 상태에 있더라도 성능이 나쁘면 서비스가 가용하지 않다고 생각한다. 또한, 그 측정에 있어서도 타임아웃과 동시에 많은 서비스 요청이 있을 경우에 안정성 측면의 품질 요소는 떨어지게 된다. 그러므로, 웹 서버에서 우선순위 할당하는 요인으로 이러한 요인들을 감안함으로써 안정성 측면의 품질 요소에 대한 차별화 서비스에 영향을 미칠 수 있다.

- 가용성(Availability) : 서비스가 존재하는지 또는 즉시 사용 가능한 상태인지의 측면에서의 품질, 즉 서비스가 이용가능한지를 나타내는 정도를 의미한다.
- 접근성(Accessability) : 서비스 요청에 대하여 서비스가 가능한 정도, 즉 특정 한 시점에서 성공적으로 서비스를 사용할 수 있는지를 의미한다. 예로 어떤 경우의 웹 서비스에 대하여 가용성은 높지만 동시에 많은 요청이 있을 경우에 잘 처리되지 않는 경우가 있다. 따라서 접근성을 보장하기 위해서 동시 사용자를 보다 많이 처리할 수 있어야 한다.
- 신뢰성(Reliability) : 서비스의 신뢰성은 일정 시간 내에 요청에 대한 응답을 보내오는 확률을 의미하며, 이는 일별, 주별, 월별, 또는 년별로 나타낼 수 있다.

위와 같은 서비스 품질 요소들을 기반으로 웹 서비스의 성능을 정의된 수준으로 보증하기 위한 서비스 제공자와 사용자간의 공식적인 계약이 필요하다. 이와 같은 서비스 수준 계약은 포괄적일 수도 있고 매우 상세화될 수도 있다. 어떤 경우 고객은 개별화된 서비스 수준 계약을 통해 서비스 제공자가 보증한 특정 성능 수준을 기대하기도 하며, 계약 이행에 실패한 경우 서비스 제공자와 사용자가 취해야 하는 단계들을 포함할 수 있다[1][3].

3. 웹 서비스 스케줄링 기법

차별화된 웹 서비스를 지원하기 위한 웹 서버의 모델은 [그림 1]과 같다. 이 모델에서는 기본적으로 웹 서비스 사용자와 제공자간에 서비스 수준 계약이 체결되어 있어야 하며, 2장에서 언급한 웹 서비스 품질 정보에 대한 측정 값이 있어야 한다. 웹 서비스 사용자들에 의하여 요청된 웹 서비스 메시지는 분류 정책에 의하여 분류되며, 실행 큐에서는 최적의 처리를 지원할 수 있는 메시지에 한하여 처리할 수 있게 한다. 그 이상의 요청이 발생하였을 경우에는 각각의 해당 버퍼 큐에서 대기하게 되며, 분류된 버퍼 큐에서 실행 큐로 보내어질 메시지는 우선순위 정책에 의하여 결정되어진다.



[그림 1] 차별화 웹 서버의 모델

이 모델에서는 먼저 웹 서버가 처리할 수 있는 최적의 메시지 수를 제어할 수 있어야 한다. 웹 서버에 요청된 메시지가 증가하게 되면 그 처리량은 포화 지점까지 다다르게 되고 응답시간 또한 현저하게 늘어나게 된다. 이는 포화 지점을 지나게되면 성능은 리소스 충돌, 컨텍스트 변환 등으로 감소하게 되기 때문이다[1]. 그러므로 웹 서비스의 개시나 서비스 수준 계약 이전에 이러한 포화 지점을 알아내야 하며, 이를 이용하여 실행 큐에 적합한 메시지의 수를 정할 수 있어야 한다.

차별화된 웹 서비스를 지원하는 웹 서버를 구축하기 위하여 크게 각 메시지를 분류하는 모듈과 분류된 메시지들에 대하여 우선순위를 할당할 수 있는 모듈이 필요

하다. 각 메시지를 분류하는 모듈은 서비스 수준 계약에 의하여 분류할 수 있으며, 우선순위를 할당하는 모듈은 축적된 웹 서비스 품질 정보를 이용하여 적용할 수 있다.

우선순위를 할당하는 모듈에서는 실행 큐가 포화상태가 되었을때 각 분류된 큐에서 대기하고 있는 메시지들에 대하여 우선순위를 할당하여 우선순위가 가장 높은 메시지를 차례대로 실행 큐로 전달하여야 한다. 본 논문에서는 이러한 우선순위를 책정하기 위하여 2장에서 도출한 웹 서비스 품질 요소를 모니터링한 값과 서비스 수준 계약에 의하여 체결된 품질 요소 값을 비교하여 전체적으로 서비스 수준 계약을 높은 수준에서 만족할 수 있도록 한다. 기본적으로 분류된 큐에서는 FIFO 방식이 적용되어 가장 먼저 들어온 메시지가 그 큐에서는 가장 먼저 서비스가 되어진다. 분류된 큐들에서 대기하고 있는 메시지들이 있을 때 각 큐에서 가장 먼저 들어온 메시지들에 대하여 비교 대상이 선정되며, 이 메시지들에 대하여 품질 정보 값을 산출하여 이 값이 가장 작은 메시지에게 가장 높은 우선순위가 할당된다.

응답 시간은 사용자가 품질 요소를 평가할 때 직접적으로 느낄 수 있는 요소이며, 서비스 수준 계약에서도 높은 비중을 두고 계약을 체결하는 경우가 많다. 이 응답 시간을 적용하기 위하여 다음과 같은 (1)의 식을 적용하여 값을 산출할 수 있다.

$$V_{RT} = RT_{SLA} - QT - RT_{mean} \quad (1)$$

(1)의 식에서 RT_{SLA}는 서비스 수준 계약에 서로 상호간에 계약된 응답 시간의 값이며, QT_{mean}는 큐에서 그 메시지가 대기한 시간 RT는 그동안 측정된 평균 응답 시간을 나타낸다. 기본적으로 (1)의 식을 이용하여 V_{RT}값이 가장 적은 즉, 만족해야 할 응답시간의 여유가 가장 조금 남아 있는 메시지에게 우선순위를 결정할 수 있다. 하지만, 본 논문에서는 응답 시간 이외에도 다른 품질 요소를 고려하여 응답 시간의 만족률과 함께 다른 품질 요소에 대한 전체적인 만족률을 향상시키기 위하여 해당 품질 요소의 만족률을 계산한다. 다음 (2)의 식은 해당 메시지의 처리량에 대한 만족률이다.

$$V_T = T / T_{SLA} \quad (2)$$

(2)의 식에서 T_{SLA}는 서비스 수준 계약에 서로 상호간에 계약된 처리량의 값이며, T는 그동안 측정된 처리량을 나타낸다. 또한, 가용성, 접근성, 그리고 신뢰성에 대한 만족률을 다음과 같이 구할 수 있다.

$$V_{Av} = Av / A_{vSLA} \quad (3)$$

$$V_{Ac} = Ac / A_{cSLA} \quad (4)$$

$$V_R = R / R_{SLA} \quad (5)$$

위의 식에서 A_{vSLA}, A_{cSLA}, 그리고 R_{SLA}는 서비스 수준 계약에 서로 상호간에 계약된 각각 가용성, 접근성, 그리고 신뢰성의 값이며, Av, Ac, 그리고 R은 그동안 측정된 가용성, 접근성, 그리고 신뢰성의 값을 나타낸다.

(1) ~ (5)의 식에서 구해진 값들을 이용하여 우선순위에 할당하는데 사용하는 값을 다음과 같이 계산할 수 있다.

$$V_P = V_{RT} * [(w_t*V_T) + (w_{av}*V_{Av}) + (w_{ac}*V_{Ac}) + (w_r*V_R)] / 4 \quad (6)$$

각 대기 큐에서 대기하고 있는 가장 먼저들어온 메시지들에 대하여 (6)의 식을 이용하여 V_P의 값을 계산 및 비교하여 가장 값이 작은 메시지에게 높은 우선순위를 부여할 수 있다. 이 때 각 품질 요소별로 가중치를 두어 사용자가 더 중요하게 생각하고 있는 품질 요소에 대하여 많은 비중을 차지하게 할 수 있다. 예를 들어, 처리량, 가용성, 접근성, 그리고 신뢰성에 대하여 같은 비중으로 서비스 수준 계약이 체결되었다면 각각의 가중치를 1로 줄 수 있다.

4. 결론 및 향후 과제

본 논문에서는 웹 서비스에서 제공자와 사용자간의 서비스 수준 계약을 만족하는 차별화 서비스를 제공하기 위하여 웹 서버 수준에서 사용할 수 있는 스케줄링 기법을 제안하였다. 이를 위하여 먼저 사용하게 되는 웹 서비스의 품질 평가 요소를 도출하고 이를 기반으로 하는 축적된 웹 서비스 품질 정보와 서비스 수준 계약에 명시된 정보를 이용하였다. 이러한 웹 서비스 품질 정보를 스케줄링에 반영함으로써 모든 서비스 수준 계약을 최대한 만족하면서 차별화 서비스를 제공하고, 우선순위가 낮은 요청된 메시지에 대한 기근(starvation) 문제를 해결함은 물론, 각 품질 요소에 대하여 가중치를 부여할 수 있게 된다.

향후 연구로는 제안한 스케줄링 기법을 활용하여 차별화 서비스를 지원하는 웹 서버를 설계하고 구현하여 다른 스케줄링 기법과 비교하여 이에 대한 효율성을 검증하도록 한다.

참고문헌

[1] Asit Dan, Heiko Ludwig, and Giovanni Pacifici, "Web services differentiation with service level agreements", White Paper, IBM, 2003.5

[2] Daniel A. Menasce, "QoS Issues in Web Services", IEEE Internet Computing, 2002.12

[3] 한국전산원, "웹 서비스 품질관리 동향 및 도입전략 연구", 2003.12

[4] S. Blake, D. Black, M. Carlson, E. Davies, Wang Z. and W. Weiss.m "An architecture for differentiated services", IETF RFC 2475, 1998.12

[5] R. Bhatti and R. Friedrich., "Web Server Support for Tiered Services", IEEE Network, 13(5):64-71, 1999.9

[6] N. Vasiliou and H. Lutfiyya, "Providing a Differentiated Quality of Service in a World Wide Web Server", Proc. Of the Performance and Architecture of Web Servers Workshop, Santa Clara, California USA, pp. 14-20, 2000.6

[7] Xiaobo Zhou, Yu Cai, Ganesh K. Godavari, and C. Edward Chow, "An Adaptive Process Allocation Strategy for Proportional Responsiveness Differentiation on Web Servers", IEEE International Conference on Web Services(ICWS 2004), pp. 142-149, 2004.7