사용자 평등성을 제공하는 온라인 웹 서비스 환경의 폭주하는 태스크 관리 기법

조홍래
고려대학교 컴퓨터정보통신대학원
chorr@hanmail.net

Congested Task Management Providing Users' Fairness On Online Web Service Environment

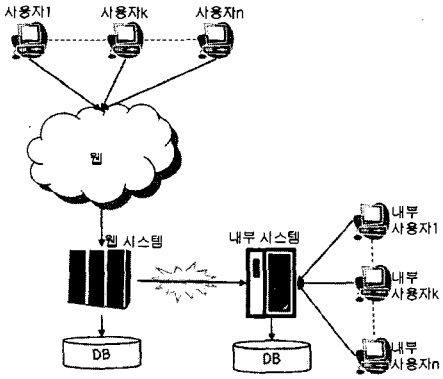
Hong Rae Cho
Graduate of Computer & Information Technology, Korea University

요 약

인터넷 사용자의 급격한 증가와 기업들의 서비스 통합으로 다양한 시스템으로 구성된 온라인 웹 환경의 서비스가 제공되고 있다. 시간제한성의 서비스를 사용할 경우 예측하지 못한 부하발생으로 시스템의 폭주 상태가 발생하게 된다. 예측하지 못한 부하에 대해 서비스 품질을 보장할 수 있는 부하제한 방법이 연구되었으나 웹 사용자의 사용패턴을 고려하지 못하고 있다. 본 논문에서는 폭주하는 서비스 상태에서 연계된 시스템 간 부하균형 상태를 유지하면서 사용자에게 동등한 만족도의 서비스 품질을 제공하는 방안을 제안하고 모의실험을 통해 서비스 성공률과 사용 평등성 비율을 분석하였다.

1. 서 론

초고속 인터넷 망의 대중적인 보급으로 인하여 인터넷 사용자는 급격히 증가하였다. 특정 시간대에 사용자가 집중되는 현상으로 응답속도에 중요한 이슈가 되었다[1]. 응답속도에 문제로 사용자들에게 불편을 주는 사례가 여러 차례 발생되었고, 이러한 사례가 발생될 경우 일반적인 웹 서버 허용 제어 장치에 의하여 요청이 차단되고 차단된 사용자는 재시도 하는 특성을 갖는다. 재시도하는 요청은 온라인 웹 서비스 환경의 시스템에 과부하를 가중시킬 확률이 매우 크다[2]. 부하의 과중을 해결하기 위하여 클러스터 환경을 구성하여 네트워크, 데이터베이스, 응용 시스템들의 확장성, 고가용성, 경제성, 신뢰성을 높이는 고성능 시스템으로 확장하는 연구가 활발히 진행되고 있다[3][4][5]. 인터넷의 급속한 보급으로 기업들의 비즈니스 환경이 빠르게 변화하고 있으며 새로운 서비스들을 경쟁적으로 도입하고 있다. 웹 서비스는 기존의 서비스들을 통합하여 새로운 서비스를 제공하고 있으면 여러 시스템에 분산되어 있거나 혹은 이기종의 시스템 환경이 통합되어 제공되고 있다[6][7]. 급속히 늘어나는 사용자와 비즈니스 업종에 따라서 다양해지는 서비스들로 인하여 서비스를 제공하는 시스템들에게 발생하는 부하는 커지고 그 특성도 다양해지고 있다. 실시간 응답을 요하는 분산 환경의 서비스는 사용자들의 부하를 예상하지 못하게 되는 경우가 발생한다[8]. [3][4][5]와 같은 클러스터 환경을 구성하여 예상된 과부하를 처리할 수 있다. 그러나, 서비스가 다양화됨에 따라 그림 1과 같이 시간제약을 갖는 서비스를 제공하는 분산 실시간 웹 환경에서 과도한 부하에 따른 서비스 지연상태일 경우에 [2]와 같이 사용자가 서비스 요청을 재시도하게 된다. 제한시간을 준수하려는 사용자의 부정확하고 예측하기 어려운 부하가 지속될 경우 서비스 제공 불능상태가 발생한다. 이러한 문제를 해결하기 위해서는 예측할 수 없는 부하를 제어하여 내부 시스템의 자원이 허용하는 만큼의 서비스를 제공해야 한다. 웹 사용자들의 사용패턴을 고려하여 반복 요청되는 부하를 효과적으로 분배할 수 있는 방안이 필요하다.



[그림 1] 폭주하는 온라인 웹 서비스 환경

본 논문에서는 이러한 이슈들에 대한 해결방안으로서 시스템이 허용하는 자원에 한하여 사용자에게 동등한 서비스 기회를 제공하는 관리기법을 제안하고자 한다. 이 정책의 목적은 복잡한 온라인 웹 환경에서 마감시한에 성격을 갖는 서비스를 제공할 경우 사용자 요청의 폭주 상태에서도 동등한 기회의 제공으로 사용자별 서비스 성공률을 향상시키기 위함이다.

2. 관련연구

본 논문은 온라인 웹 서비스 환경에서의 부정확하고 예측할 수 없는 폭주하는 부하에 대해 사용자에게 동등한 서비스를 제공하는 정책과 시스템 가용자원 제공을 위한 태스크 분배정책과 관련된다.

2.1 사용자 평등성 제어 모델에 대한 연구

제한된 시스템 자원을 제공하는 환경에서 안정된 서비스 품질을 유지하면서 사용자의 서비스 요구를 제어하는 기술로 차별화된 부하제어 기법이 연구되었다. [9]에서는 기업의 수익성에 기초하여 고객을 세분화하였고 콘텐츠 서비스의 CRM 기반을 마련하기 위해 수익성이 높은 고객에게 서버자원을 이용할 기회를 더 많이 부여한다. 서버 자원을 할당하고 고객 등급별 수락률을 제어하기 위한 알고리즘으로 CRFA(Client Request Filtering Algorithm)을 제시하고 있다. 이 알고리즘은 서버의 가용자원을 미디어 서버로부터 제공받아 세분화된 고객에게 자원을 배부하고 있다. 미디어 서버에서 가용자원을 산출하는 부분에 대한 연구내용이 없으며 이를 보완하기 위해 예측 가능한 부하의 가용자원을 산출하는 방법을 관련연구로 참고하고 있다. 사용자에게 차별화된 부하분배를 목적으로 하는 연구이므로 본 논문에서 제시하는 이슈를 해결할 수 있는 연구인 평등성을 제공하는 방법에 적용하기에 부적합하다.

2.2 예측할 수 없는 부하제어에 관한 연구

예측할 수 없는 부정확한 부하의 계산방법(Imprecise Computations)에 대한 연구와 계산된 결과에 따라 시스템에 반영하는 방법(Feedback Control)에 대한 연구 분야로 나누어진다[8]. 부정확한 부하의 예측계산방법은 일시적인 부하가 발생되는 동안 자원 분배와 부하 제한으로 처리의 유연성을 제공한다. [8]에서는 예측 불가능한 부하의 관리 방안과 서비스 품질 요소 관리 방안을 연구하였고 실시간 데이터베이스분야에서 서비스 품질을 관리한 첫 번째 연구였다. 부정확한 부하의 계산방법과 서비스 자원 재분배 제어에 관하여 연구되었고 마감시간 초과율 및 시스템 가용성을 고려한 부하제한 알고리즘과 평균 트랜잭션 오류율을 고려한 부하제한 알고리즘으로 구분하여 제시하고 있다. [9]를 포함한 다수의 관련 논문에서 지적하였듯이 [8]에서 제안한 전체 시스템 자원의 균등한 자원 분배 기법으로 본 논문에서 제시하는 모든 이슈를 해결할 수 없다. 사용자의 도착한 순서대로 자원을 할당하는 관리기법으로 시스템을 운영하게 되면 과부하상태에서 사용자들은 성공확률을 높이기 위해 서비스를 재시도하는 특성에 따라 시스템 폭주를 발생시키게 된다. 또한 실시간 데이터베이스 자원의 가용성을 고려한 부하제한 방법이므로 부하 예측 시스템과 연계한 내부 시스템의 자원 가용성을 예측할 수 있는 방법이 필요하다.

3.사용자 평등성을 제공하는 폭주 태스크 관리기법

본 논문에서는 사용자의 과도한 부하발생에 따라 온라인 웹 서비스 환경에서 내부 시스템과의 부하 불균형으로 발생되는 사용자의 예측 불가능한 부하를 재분배하는 태스크 관리기법으로 사용자 평등성을 제공하고자 한다.

사용자에게 평등성을 제공하기 위해 웹 사용자를 구분하는 방법에 대한 연구 중 시간제약 서비스를 제공하는 온라인 웹 환경에서는 사용자정보를 사용하여 구분하는 기법을 본 논문에서 사용한다[10]. 시간제한의 특성이 있는 서비스는 회원정보를 사전에 등록하여 시간제한에 영향이 없는 형태의 서비스를 제공하므로 [10]에서 제안하는 방법 중 사용자 정보를 사용한 구분기법을 본 논문에 적합한 기법으로 채택하였다. 불확실한 부하를 제어하기 위해 웹 시스템과 내부 시스템 간 부하 불균형 시 발생하는 마감시한 초과에 대한 오류율을 반영하였고 웹 시스템에서 부하를 조절하도록 하였다. 이와 같은 기법으로 사용자에게 서비스 품질을 보장하기 위한 알고리즘은 그림 2와 같다.

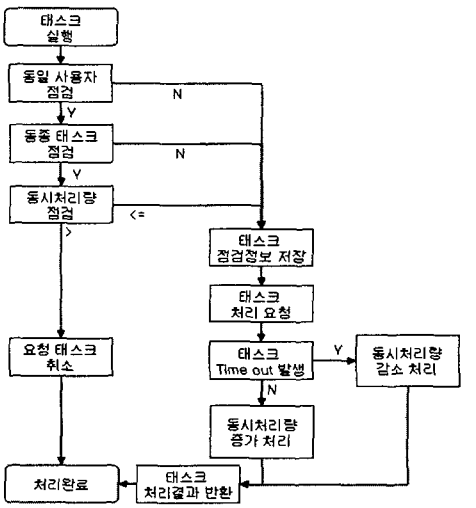


그림 2. 사용자 평등성을 제공하는 폭주 태스크 분배 알고리즘

그림 2의 동일 사용자 및 동종 태스크의 요청이 있을 경우 현재의 동시 처리량을 확인하여 과부하 요청일 경우에는 요청 태스크를 취소시킨다. 병목구간이 발생되는 내부 시스템간의 연계부분에서 마감시간을 초과하는 태스크가 발생되면 요청되는 태스크의 부하를 조절하여 내부 시스템간의 부하 균형을 맞추게 된다.

본 논문의 제안 기법은 서비스별로 사용자에게 현재 가용한 시스템 자원을 공평하게 할당하는 기법이다. a ~ z의 사용자가 요청하는 상대적 요청비율[11]을 $F = \sum_{u=a}^{z} F_u$ = 1라고 가정하자. 과부하시점 k의 사용자별 오차율 편차 $sdte(k)$는 다음과 같다.

$$sdte(k) = \sqrt{\frac{\sum_{u=a}^{z}(\min(tte(k)\times\frac{F_u}{F}, tte(k)\times\frac{1}{z}) - ate(k))^2}{|term(k)|-1}} \quad \text{(식1)}$$

- 총 처리 태스크 수 $term(k)$
- 전체 사용자 태스크 오류율 $tte(k) = \frac{\gamma(k)\times\delta l(k)\times z}{term(k)}$
- 사용자별 요청 허용 수 $\delta l(k)$
- 요청률과 처리오류율 간 비례값 γ, $(0<\gamma<1)$
- 평균 태스크 오류 ate(k) = $\frac{tte(k)}{|term(k)|}$

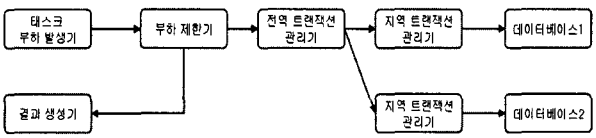
(식1)의 $\min(tte(k)\times\frac{F_u}{F}, tte(k)\times\frac{1}{z})$은 부하가 증가할수록 $tte(k)\times\frac{F_u}{F}$가 커지므로 $tte(k)\times\frac{1}{z}$을 사용하는 비율이 높아진다. 따라서 부하가 증가할수록 사용자별 오차율 편차가 감소하게 됨으로써 사용자에게 공평한 태스크 처리기회를 제공할 수 있다.

4. 실험

본 논문에서는 제안한 알고리즘의 사용자 평등성을 평가하기 위해서 다수 사용자의 동시 접속수를 변화시켜 태스크 성공률 및 사용자별 성공률 편차에 대한 시뮬레이션을 수행하였다. 다음과 같이 실험환경과 실험결과를 분석한다.

4.1 실험 방법

[8]에서는 비정형적이며 예측하기 어려운 부하에 대한 부하 재분배 알고리즘의 성능실험이 있었다. 본 논문은 [8]에서 제안한 알고리즘에서 지원하지 못하는 사용자 평등성을 본 논문의 알고리즘이 제공하고 있다. 제안하는 알고리즘이 [8]의 알고리즘보다 사용자 평등성의 개선된 결과를 실험하고자 하였다. 실험 시뮬레이션 환경은 2대의 PC에서 J2EE기반의 트랜잭션 서버를 사용하여 두 개의 데이터베이스의 데이터를 처리하도록 구성하였으며 시뮬레이션 시스템 구조는 다음과 같다.



[그림 3] 시뮬레이션 모델

부하 제한기 부분의 부하 재분배 알고리즘을 적용하여 사용자 평등성을 결과 생성기로 전송하였다. 태스크 부하 발생기에서 생성하는 입력 파라미터는 표1과 같이 설정하였다.

표 1. 시뮬레이션 입력 파라미터

항목	설정 값
마감시한	20초
부하지속시간	5분
사용자수	5 ~ 30명
사용자당 요청수	5 ~ 30회
요청주기	10 ~ 1000ms 비정형적 발생

4.2 실험 결과

사용자수와 사용자당 요청수를 증가시킴으로써 시뮬레이션 시스템에 부하를 증가시켜 실험결과를 측정하였다.

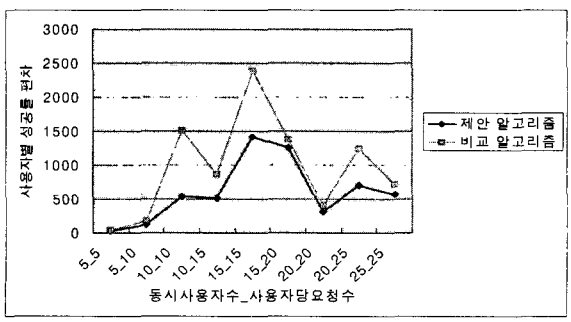


그림 4. 요청부하별 태스크 성공률 편차 비교

전체 성공률은 제안 알고리즘이 다소 우수하게 측정되었다. 비교 알고리즘보다 사용자의 트랜잭션 별로 부하를 제한하기 때문에 성공률이 높은 것으로 평가된다. 그림 4에서와 같이 사용자별 성공률 편차를 측정한 결과는 제안 알고리즘의 편차가 작으므로 비교 알고리즘에 비해 사용자에게 공정한 기회를 제공함을 알 수 있었다.

5. 결론

제안하는 알고리즘은 웹 시스템과 내부 시스템간의 부하 불균형을 해소하고 과부하 태스크를 관리하는 기법을 제공하였다. 사용자별로 태스크 성공률 편차를 감소시켜 과부하 상태에서도 전체 사용자에 대해 만족도를 향상시켜 서비스 품질을 유지하는 결과를 얻었다. 과부하 상태에서 사용자의 재시도 요청으로 시스템 폭주상태가 발생되지 않도록 부하를 제한하면서 사용자 평등성을 유지한다면 안정적인 서비스 품질을 제공하는 시스템을 운영할 수 있다.

향후 서비스별 과부하 발생 시 타 서비스의 영향 및 부하제한기가 운영중인 시스템의 자원 영향에 따른 부하 제한기법이 연구되어야 할 것이다.

6. 참고문헌

[1] 노경혁 외 1명, "웹사이트의 사용자 응답속도 향상을 위한 연구", 한양대학교 산업대학교 석사학위 논문, 2004

[2] 이철 외 3명, "사용자 평등성을 제공하는 웹 서버 허용 장치 제어 장치 설계", 한국정보과학회, VOL. 29. NO. 2, pp 538~540, 2002

[3] 박혜숙 외 2명, "액세스 망에서의 DiffServ기반 가입자 대역 보장 방법 연구", 정보처리학회, VOL. 12-C. NO. 5, pp 709~716, 2005

[4] 황미선 외 2명, "차별화 서비스를 위한 이질 웹 서버 승인 제어", 정보처리학회, VOL. 12. NO. 1, pp 1509~1512, 2005

[5] Stephane Gancarski et al, "Parrallel Processing with Autonomous Databases in Cluster Systems", In 10th Int. Conf. on Cooperative Information Systems, LNCS2019, pp 410~428, 2002

[6] R. Akkiraju, et al., "A Framework for Facilitating Dynamic e-Business Via Web Services", OOPLSA2001, Florida, USA, 2001

[7] S. Tsur, "Are Web Services the Next Revolution in E-commerce?", Proc. of the 27th VLDB, Roma, Italy, 2001

[8] M. Amirijoo, et al., "Specification and Management of QoS in Real-Time Database Supporting Imprecise Computations", IEEE Transactions on Computers, VOL. 55. NO. 3, pp 304~319, 2006

[9] 박혜숙 외 1명, "군집분석(Cluster Analysis)을 활용한 사용자 등급 기반의 서비스 수락 정책", 정보처리학회 VOL. 12. NO. 3, pp 461~470, 2004

[10] 최영환 외 1명, "웹 마이닝을 위한 입력 데이터의 전처리 과정에서 사용자구분과 세션정보", 정보과학회 VOL 30. NO. 9, pp 843~849, 2003

[11] 박재석 외 1명, "데이터 스트림에 대한 연속 질의를 위한 우선순위 기반의 의미적 부하제한", 데이터베이스연구회 KDBC VOL. 21, NO. 01, pp 1~7, 2005