강화신호를 이용한 건물공조시스템의 최적제어에 관한 연구

조 성 환†, 양 성 희*, 양 훈 철
한국에너지기술연구원 건물에너지연구센터, *한양대학교 건축공학과

A Study of Optimum Control in Building HVAC System using Reinforce Signal

Sung-Hwan Cho†, Sung-Hee Yang*, Hooncheul Yang
Building Energy Research Center, KIER, Yuseong-gu Taejeon 305-343, Korea
**Department of Architecture Engineering, Hanyang University, Seoul 133-791, Korea*

(Received April 21 2004; revision received September 22, 2004)

ABSTRACT: Technology on the proportional integral (PI) control have grown rapidly owing to the needs for the robust capacity of the controllers from industrial building sectors. However, PI controller requires tuning of gains for optimal control when the outside weather condition changes. The present study presents the possibility of reinforcement learning (RL) control algorithm with PI controller adapted in the HVAC system. The optimal design criteria of RL controller was proposed in the Environment Chamber experiment and a theoretical analysis was also conducted using TRNSYS program.

Key words: Reinforcement learning(강화학습), Optimal control(최적제어), Auto-tuning(자동동조), Building automation system(건물자동화 시스템), PI control(비례적분제어)

기 호 설 명

a : 행동
E : 확률
K : 제어게인
Q : 목표값
R : 강화신호
s : 상태
T : 온도 [℃]

그리스 문자

α : 학습률
γ : 감쇠계수

하첨자

i : 적분제어
p : 비례제어
t : 시간
π : 정책

† Corresponding author
Tel.: +82-42-860-3236; fax: +82-42-860-3202
E-mail address: shcho@kier.re.kr

1. 서 론

현재 공기조화 및 냉동기기 등 대부분의 산업용 제어기에는 PI(Proportional-Integral) 제어기를 사용하고 있으며 운전환경이 변화하는 경우에는 최적의 제어성능을 유지하기가 어렵고 이에 따른 부적절한 제어는 에너지 소비량을 증가시킨다.[1] 따라서 적절한 동조(tuning)를 통해서 플랜

트의 동특성을 추출할 필요가 있다. 그러나 동조 과정은 많은 시간과 경비가 소모될 뿐만 아니라 강한 비선형이나 큰 지연시간을 갖는 시스템에서는 적용이 어렵다. 또한 동조 후에도 제어성능이 변화될 수 있으므로 최적의 제어성능을 유지하기 위해서는 재동조과정이 필요하다.[2,3]

이를 해결하기 위한 방법으로 자기동조기능을 갖는 제어 알고리즘인 신경망(neural network) 제어와 강화학습(reinforcement learning, RL) 제어 등이 알려져 있다. 신경망 제어는 복잡한 회로망에서는 학습속도가 느리고, 뉴런의 포화영역에서는 동작하지 않으며 오프라인 상태에서 기존 데이터를 충분히 확보해야 하는 단점이 있다.

강화학습에 대한 기존의 연구에서 Watkins et al.[4]은 최적학습방법으로 Q-학습법(learning)을 제안하였다. Anderson et al.[5,6]은 난방코일 시뮬레이션에 강화학습제어를 적용하여 제어변수들에 대한 이론과 실험값 설정, 학습에 따른 실행결과를 신경망 제어, PI 제어와 함께 비교하였다. Barto et al.[7]은 실시간 동특성을 가진 프로그램의 학습과 실행을 연구하였다. Sutton[8,9]은 강화학습제어의 평가함수학습인 Temporal Difference(TD) 방법을 사용하였다.

본 연구는 강화학습제어의 특성을 규명코자 온라인 상태에서 학습제어가 가능하고 자기동조기능을 보유한 방법으로 PI 제어기의 출력제어신호를 보상하여 주는 강화학습 제어 알고리즘 방법을 사용하여 HVAC 시스템의 제어성능을 개선시킬 수 있는 방법을 제안하였다. 이를 위해 실제 건물의 공조시스템에 제어방법을 적용하여 실험적으로 연구하였다. 또한 이론적인 연구방법으로 TRNSYS 프로그램의 한 모듈인 제어 알고리즘(PI, RL)을 개발하여 다양한 조건들의 변화에 대해 동적 시뮬레이션을 수행함으로써 강화학습제어기에서 가장 중요한 인자 중의 하나인 보상제어방법의 최적설계방법을 제안하였다.

2. 제어 알고리즘

2.1 강화학습 알고리즘

강화학습은 환경에 대한 충분한 지식이 없어도 강화신호에 의하여 주어진 환경에 따른 시행착오를 통해서 최적의 행동을 학습할 수 있는 방법이다. Watkins et al.[3]의 Q-학습법은 가장 널리 사용되는 강화학습방법이다. Q-학습법은 미래의 행위에 대한 보상에 감쇠를 고려한 최적의 행위를 찾아내는 알고리즘이다. 이 알고리즘은 상태-행위의 조합으로 이루어진 행위 값(Q-value)을 정의하여 각 상태에서의 최적의 행위를 수행하는 기법이다.

강화학습은 두 가지의 구성요소를 통해서 학습이 이루어진다. 하나는 행위자(agent)이고, 다른 하나는 환경(environment)이다. 행위자는 환경에서 주어진 상태에 따라 적절한 행위(action)를 환경에 행하며, 환경은 주어진 행위에 따라 변화된 상태와 행위가 적절했는가에 대한 판단을 강화신호로 행위자에게 보내는데, 이와 같은 과정을 반복하며 학습이 이루어지게 된다.

기본적인 Q-학습법[4]은 다음과 같다.

(1) 상태, s와 행위, a에 대한 목표값, $Q(S,a)$를 임의의 값(일반적으로 0)으로 초기화한다.

(2) 현재의 상태, s를 인식한다.

(3) 상태-행위 규칙에 따라 행위, a를 택한다.

(4) 주어진 상태에서 행위, a를 수행하고 다음 환경을 s_t라 한다.

(5) 강화의 결과로서 적용된 행위, a_t와 그때의 상태, s_t에서 보상을 $R(s_t, a_t)$로 정의한다.

(6) 목표값, $Q_\pi(s_t, a_t)$는 주어진 상태, s_t와 행위, a_t에서 평가함수(value function)이며 식(1)과 같이 표현된다.

$$Q_\pi(s_t, a_t) = E_\pi\left\{\sum_{k=0}^{T}\gamma^k R(s_{t+k}, a_{t+k})\right\} \quad (1)$$

여기서 γ는 0과 1의 사이의 감쇠계수이다.

(7) 식(1)은 강화 $R(s_t, a_t)$를 더하여 순시보상(immediate reinforcement)과 미래 보상(future reinforcement)의 합을 더한 값으로 다시 표현할 수 있다.

$$\begin{aligned} Q_\pi(s_t, a_t) &= E_\pi\left\{R(s_t, a_t) + \sum_{k=1}^{T}\gamma^k R(s_{t+k}, a_{t+k})\right\} \quad (2)\\ &= E_\pi\left\{R(s_t, a_t) + \gamma\sum_{k=0}^{T-1}\gamma^k R(s_{t+k+1}, a_{t+k+1})\right\} \end{aligned}$$

(8) 동적 프로그램에서 정책평가(policy evaluation)는 평가함수, Q_π가 원하는 합산값에 수렴할 때까지 반복적으로 계산하여 얻어진다. 반복적 정책평가방법은 식(3)과 같이 평가함수, Q_π의 현재값으로 나타난다.

$$\triangle Q_\pi(s_t, a_t) = E_\pi\{R(s_t, a_t) + \gamma Q_\pi(s_{t+1}, a_{t+1})\} - Q_\pi(s_t, a_t) \quad (3)$$

여기서 기대값은 가능한 다음 상태, s_{t+1}에 대해 얻어진다. 기대값은 상태변환함수(state transition probability)에서 하나의 모델을 요구하며, 만약 모델이 없다면 Monte Carlo 방법이 사용된다.

(9) 행동-선택 정책을 향상시키고 최적 제어를 달성하기 위해 평가반복(value iteration) 프로그램 기법이 적용되며, 이것은 정책평가와 정책개선을 조합한 것이다.

$$\triangle Q_\pi(s_t, a_t) = a_t[R(s_t, a_t) + \gamma Q_\pi(s_{t+1}, a_{t+1}) - Q_\pi(s_t, a_t)] \quad (4)$$

여기서 학습률, a_t의 범위는 $0 \leq a_t \leq 1$이다.

(10) 총 보상값을 최소화하는 가정 하에서 목표값, Q에 대한 평가반복의 Monte Carlo식 표현법은 식(5)와 같다.

$$\triangle Q_\pi(s_t, a_t) = a_t\Big[R(s_t, a_t) + \gamma \min_{a' \in A} Q_\pi(s_{t+1}, a') - Q_\pi(s_t, a_t)\Big] \quad (5)$$

이것이 Q-학습법 알고리즘이며 Watkins et al.[3]은 행위, a를 선택할 때 상태, s에 대해 목표값, Q를 최소화하는 것이 최적의 보상의 합이 됨을 증명하였다. Fig. 1은 PI 제어기와 결합된 RL 제어기의 구조도를 나타낸다.

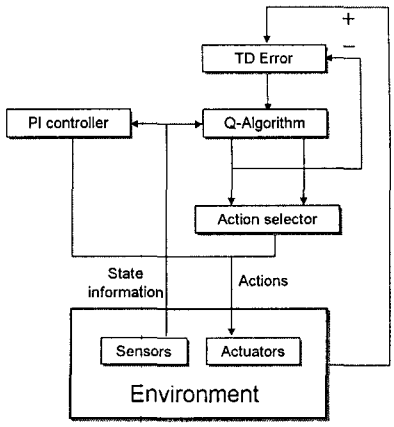


Fig. 1 Structure of RL controller combined with PI controller.

3. 실험장치 및 결과

실제 건물의 공조시스템에 강화학습 제어방식의 적용 가능 여부를 검토하기 위하여 한국에너지기술연구원에 소재한 인공기후실험동(environment chamber)에서 동절기 조건을 재연시킨 후 PI 제어, 강화학습 제어기의 특성을 실험하였다. 이때 RL 제어기는 기본적으로 PI 제어기를 기반으로 하여 제어게인(K_p, K_i)을 강화시켜 줄 수 있게 구성되어 있다.

3.1 인공기후실험동

인공기후실험동은 실내 환경, 에너지소비량, 설

Table 1 Operation range of experiment system

	Operation range
Indoor condition	24℃ (75.2°F)
Outdoor condition	−5~10℃
Supply fan	Max : 1000 CMH (0.278 m^3/s), Min : 200 CMH (0.055 m^3/s)
Return fan	Max : 900 CMH (0.250 m^3/s), Min : 200 CMH (0.055 m^3/s)
Cooling coil	Capacity : 13,608 kcal/h (4.5 HP), Condenser : 9,072 kcal/h (3.0 HP), 4,536 kcal/h (1.5 HP), Inlet cooling water temp : 7℃, Outlet cooling water temp : 13℃
Supply set pressure	45 mmAq (448 Pa)

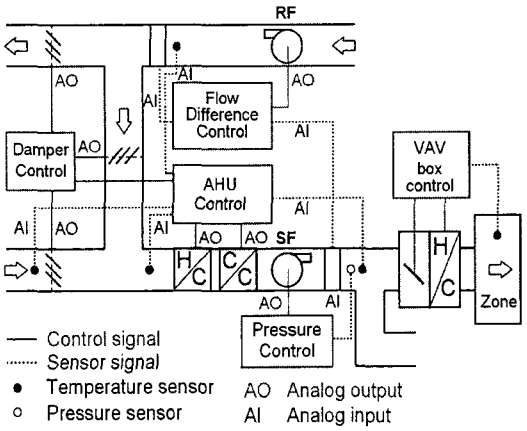


Fig. 2 Schematic diagram of AHU.

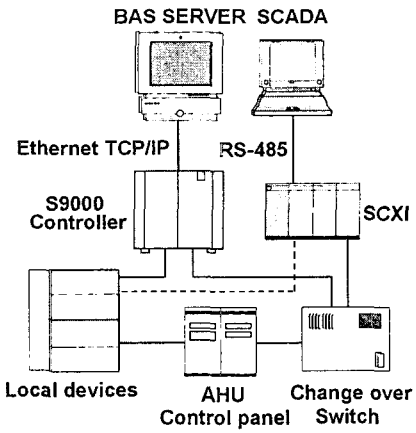


Fig. 3 Schematic diagram of overall system.

비용량 등 건물 전체의 열성능 해석, 설비시스템의 종합적인 성능을 분석하고 평가하기 위해 외기조건을 인공적으로 재현시켜서 연구를 수행할 수 있는 실험장비이다.

인공기후실험동 내부에 설치된 시험주택은 건물의 냉난방부하, 냉난방 설비의 효율, 열환경, 에너지절약, 건물 구조체의 열전달 현상 및 HVAC 제어에 관한 종합적인 실험을 수행할 수 있도록 구성되어 있다. 실시간 고장진단 시스템의 적용대상 건물인 인공기후실험동 내부에 있는 실험주택의 내부장비 가동조건은 Table 1과 같다.

Fig. 2는 실험주택의 비온돌 실험실과 환경 실험실에 설치된 가변풍량(VAV)방식의 공조시스템으로 실내부하 변동에 따라 급기온도를 일정하게 유지시키고 실별, 구역별 풍량을 변화시켜서 실온을 제어하고 있다.

3.2 시스템 구현

공조시스템의 감시운영제어는 주컴퓨터의 감시제어(supervisory control)와 현장제어(local loop control)로 이루어진다. Fig. 3은 시험주택의 공조시스템을 자동으로 운전하기 위해 구성한 감시운영 제어시스템을 보여준다.[10] 시스템의 구성은 기존의 감시운영 제어시스템(BAS)과 강화학습 알고리즘의 성능실험용 감시운영 제어시스템(SCADA)을 독립적으로 구성하였다. 기존의 감시운영 제어시스템에서는 주컴퓨터의 감시제어와 현장제어로 Ethernet TCP/IP를 기반으로 데이터 인터페이스를 사용하여 실시간 데이터 감시 및 운영제어를 수행하였다. 하지만 실제 다양한 제어 알고리즘에 대한 성능실험에는 한계가 있기 때문에 본 실험운영에 맞도록 PI 제어와 RL 제어 알고리즘을 적용한 감시운영제어 프로그램을 구현하여 제어기 실증실험을 통한 성능특성을 비교 분석하였다.

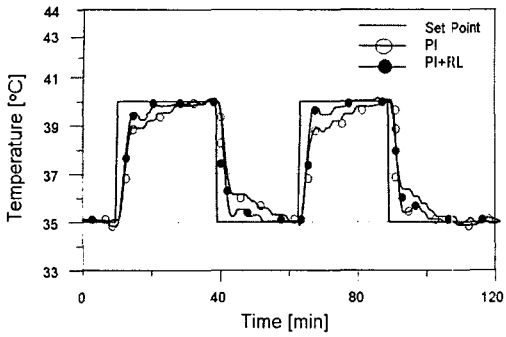
3.3 실험결과

HVAC 시스템을 위한 RL 제어기의 특성을 평가하기 위하여 겨울철 조건인 5℃로 외부조건을 설정하고 실내온도 조건은 24℃로 유지하면서 PI 제어기와 RL 제어기에 의해 난방코일을 제어하였다.

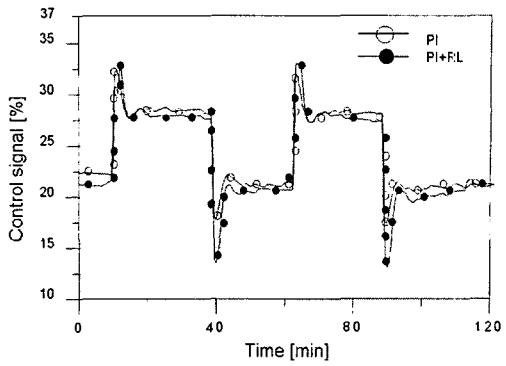
Fig. 4는 실험을 통하여 최적제어게인 K_p=1.9, K_i=7.5의 PI 제어기를 사용하는 경우와 학습이 완료된 RL 제어기를 적용하는 경우에 난방코일 제어성능 특성을 나타낸다.

Fig. 4(a)는 실내온도의 설정을 35℃에서 40℃로 단계적으로 변경시킨 경우 PI 제어와 RL 제어기에 의한 제어결과를 나타낸다.

RL 제어기를 사용하는 경우에는 Fig. 4(b)와 같이 외부환경의 변화에 따른 급기온도 설정값 변화시 PI 제어기보다 정상상태의 오차를 상당히 감소시키고, 빠른 응답성을 가지는 제어기의 성능이 전반적으로 개선되는 것을 알 수 있다. 여기서 RL 제어기는 설정조건이 변경될 때 PI 제어기보다 출력제어의 신호를 강화시켜 주는 것을 알 수 있다.



(a) Supply air temperature



(b) Control signal

Fig. 4 Control performance of PI controller and RL controller combined with PI controller.

4. 설계인자의 최적화

RL 제어기의 설계시 중요한 인자 중의 하나는 제어기의 입력값과 출력값의 오차가 큰 경우에 출력값을 강화시키는 설정조건을 구하는 것이다. 본 연구는 동절기 조건에서 여러 강화신호에 따른 추종성능을 검토함으로써 RL 제어기의 설계조건을 제안하였다. 이를 위하여 TRNSYS 프로그램에 PI 제어기와 RL 제어기 부분을 Fin Efficiency와 Effectiveness-NTU를 이용한 해석방법과 복사난방공간의 온도제어방법을 프로그램화하여 TRNSYS 프로그램의 각 모듈로 이용하였다.

4.1 TRNSYS 프로그램

Fig. 5는 TRNSYS 프로그램의 계통도를 나타낸다. 실내에 공급되는 공기는 Type 32인 난방코일 내에서 가열되어지고 이는 송풍기를 통하여 대상

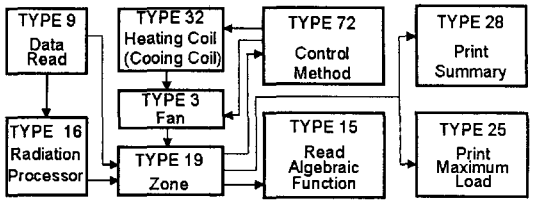


Fig. 5 Flow chart of TRNSYS program.

구역(Type 19)으로 유입된다. 구역으로 유입된 공기는 Type 72인 제어방법에 정보를 제공하고 이는 설정된 PI 제어기, RL 제어기에서 난방코일이나 냉방코일, 송풍기 등으로 제어신호를 출력시킴으로써 각 환경에 맞는 제어를 한다. Type 19 구

Table 2 TRNSYS input data

System	Parameter	Data
Heating coil (Cooling coil)	Number of row deep	6
	Number of parallel cooling circuits	24
	Coil face area (m^2)	4.17
	Inside tube diameter (m)	0.02
	Inlet air dry bulb temp (℃)	15
	Inlet air wet bulb temp (℃)	15
	Mass flowrate of air (kg/hr)	500
	Inlet water temp (℃)	10
	Mass flowrate of water (kg/hr)	100
Fan	Max flowrate (kg/hr)	2,500
	Fluid specific heat (kJ/kg-℃)	1.012
	Max power consumption (kJ/hr)	3,500
	Fraction of pump power converted to fluid thermal energy ($0<f_{par}<1$)	0.9
	Inlet fluid temp. (℃)	15
	Inlet mass flowrate (kg/hr)	100
	Control function	γ
Duct	Pipe inside diameter (m)	0.35
	Pipe length (m)	3.45
	Overall loss coefficient based on inside pipe surface area ($kJ/hr.m^2℃$)	2
	Fluid density (kg/m^3)	1.2
	Fluid specific heat (kJ/kg℃)	1.012
	Initial fluid temperature (℃)	15
	Inlet fluid temp. (℃)	20
	Inlet mass flowrate (kg/hr)	100
	Outside temperature (℃)	25

역에서 얻어진 정보는 Type 28인 출력요약부로 전달되어 환경에 대한 제어 데이터를 획득한다.

Table 2는 TRNSYS 프로그램에서 사용된 변수와 입력값을 나타낸다. 이때 입력값은 실제 대상건물에 이미 설치된 각 장비들에 대한 실제값을 기준으로 작성되었다.

4.2 PI 제어기의 최적 게인값

이론적 검토를 위하여 외기온으로 설정된 여름철 외기온도조건과 겨울철 외기온도조건은 각각 35℃와 5℃이며 공급유량은 600 kg/hr이다. Fig. 6과 Fig. 7은 PI 제어기의 게인값이 K_p=6.0, K_i=2.5인 경우, K_p=1.5, K_i=4.0인 경우 및 K_p=0.1, K_i=0.5인 경우에 실내온도가 설정온도에 대해 추종하는 결과를 나타낸다. 실내온도는 K_p=1.5, K_i=4.0에서 설정온도를 충분히 추종하지 않지만 K_p=6.0, K_i=2.5에서는 냉난방 두 조건에서 설정온도에 따라 추종하고 있다.

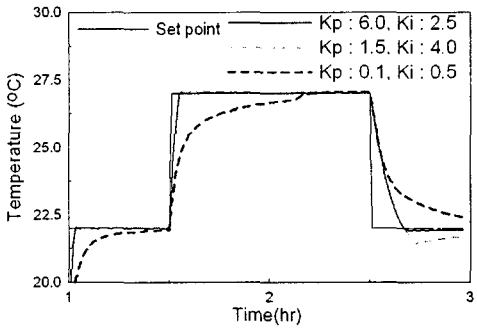


Fig. 6 Response of PI controller in heating mode (flow rate: 600 kg/hr).

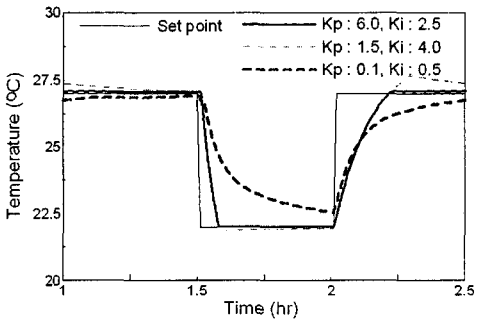
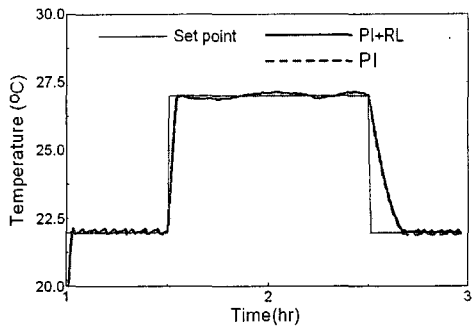


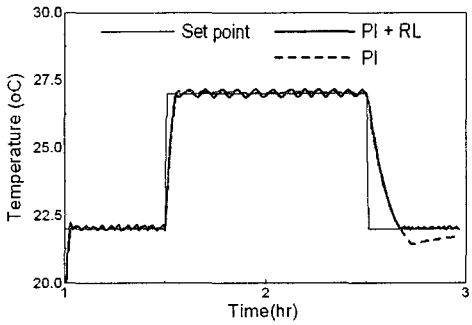
Fig. 7 Response of PI controller in cooling mode (flow rate: 600 kg/hr).

5. 강화학습의 보상신호에 따른 최적화

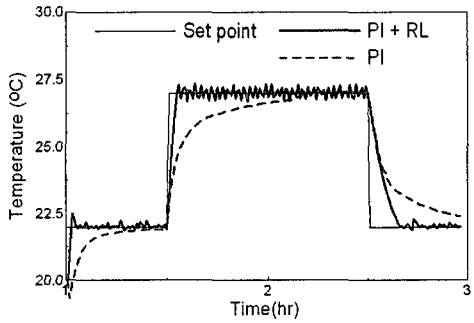
PI 제어기와 결합된 RL 제어기의 장점은 제어게인(K_p, K_i)이 적절히 조절되지 않거나 외부조건이 급격히 변화되더라도 제어신호를 강화시켜줌으로써 외란조건에 적절히 대응하여 설정값을 유지하는 것이 특성이다. 따라서 본 연구에서는 PI 제어기에서 검토된 게인값인 K_p, K_i를 변경시키면서 겨울철(외기온도 5℃)과 여름철(외기온



(a) K_p=6.0, K_i=2.5



(b) K_p=1.5, K_i=4.0

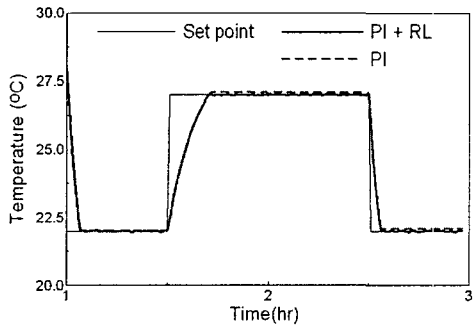


(c) K_p=0.1, K_i=0.5

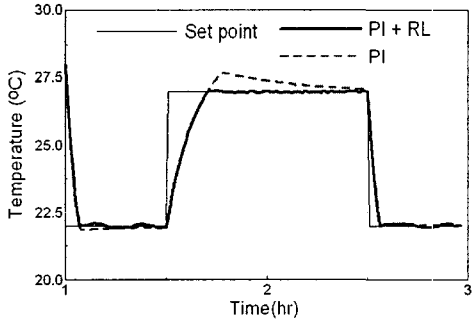
Fig. 8 Response of PI, RL controller under winter condition.

도 35℃) 조건 하에서 PI 제어기와 RL 제어기의 성능을 비교하였다.

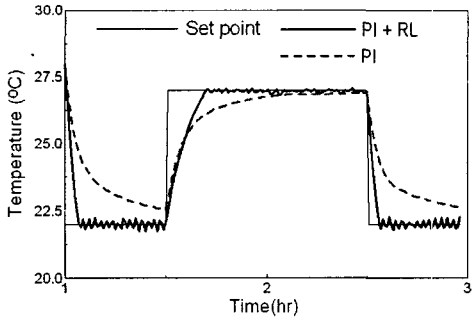
Fig. 8은 겨울철에 실내의 설정온도를 22℃와 27℃로 변화시키면서 난방코일을 통한 급기온도를 제어한 경우이다. 이때 RL 제어기의 강화값은 −0.5~+0.5로 설정하였다. 게인값이 K_p=6.0, K_i=2.5인 경우 즉 게인이 적절하여 PI 제어가 적용되는 경우에는 RL 제어기의 영향이 나타나지 않는다. 그러나 K_p=0.1, K_i=0.5로 감소된 경우



(a) K_p=6.0, K_i=2.5



(b) K_p=1.5, K_i=4.0



(c) K_p=0.1, K_i=0.5

Fig. 9 Response of PI, RL controller under summer condition.

즉 PI 제어기의 게인값이 시스템에 적절치 않은 경우에는 실내온도가 설정온도를 추종하는데 시간이 많이 소요된다. RL 제어기를 사용하면 PI 제어기의 게인이 적절히 적용된 경우와 같이 PI 제어기의 게인값이 적절치 못한 경우에도 설정온도에 대한 추종능력이 증대함을 알 수 있다.

Fig. 9는 여름철 조건(외기온도 35℃)에서 실내의 설정온도를 22℃와 27℃로 변화시키는 경우 PI 게인값의 변화에 대한 급기온도의 제어를 나타낸다. 게인값 K_p, K_i가 적절히 조정되지 못한 경우에도 RL 제어기와 함께 PI 제어기를 사용하면 설정온도에 대한 추종능력이 증대한다. 즉, RL 제어는 겨울철 난방의 경우와 마찬가지로 여름철 냉방의 경우에서도 외부환경에 적응함을 알 수 있다.

Fig. 10과 Fig. 11은 K_p=0.1, K_i=0.5인 경우, 즉 PI 제어의 게인값이 적절치 못한 경우를 대상으

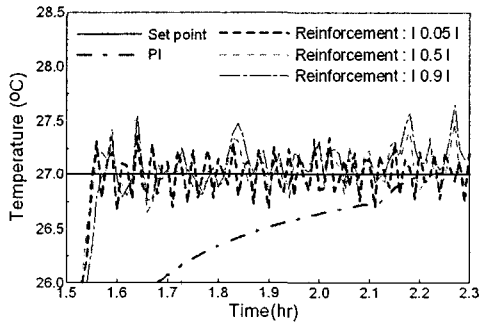


Fig. 10 Response of PI, RL controller on the variations of action ratio under winter condition (K_p=0.1, K_i=0.5).

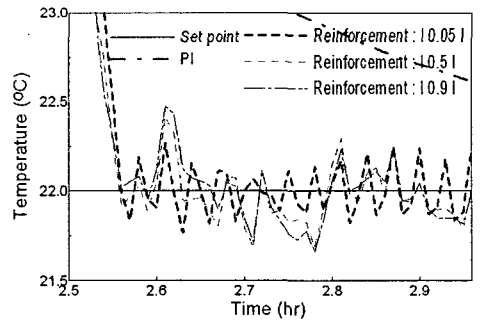


Fig. 11 Response of PI, RL controller on the variations of action ratio under summer condition (K_p=0.1, K_i=0.5).

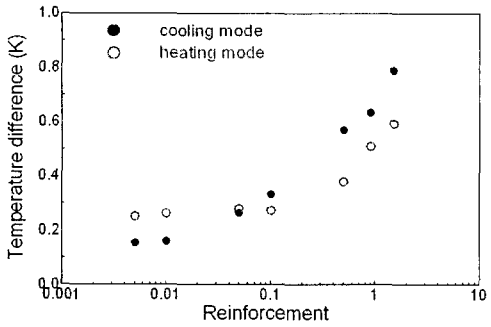


Fig. 12 Temperature difference ($T_{max} - T_{min}$) of indoor temperature with respect to reinforcement variation using PI, RL controller under summer and winter conditions (K_p=0.1, K_i=0.5).

로 여름철 및 겨울철의 외기조건하에서 RL 제어기의 강화신호를 0.05~0.90까지 변화시킨 경우이다. 겨울철 외기조건에서는 Fig. 10과 같이 강화신호가 0.05인 경우 설정온도를 추종하는 능력이 강화됨을 알 수 있다. 여름철 조건에서도 Fig. 11과 같이 강화신호가 0.05인 경우 추종능력이 좋아지는 결과를 나타내고 있다. Fig. 12는 냉방과 난방조건에서 강화신호를 0.005~1.5로 변화시키면서 실내온도의 최대값과 최소값의 차를 나타낸 것이다. 두 모드에서 강화신호가 0.05 이하에서 거의 일정한 값을 나타낸다. 따라서 실제로 RL 제어기를 설계할 경우에는 강화신호 비율을 낮게 지정하는 것이 냉난방의 제어에 유리한 결과를 나타낸다.

6. 결 론

건물공조시스템의 제어성능을 개선시킬 수 있는 방법의 하나로서 PI 제어기의 자기 동조 기능을 갖는 새로운 강화학습 제어 알고리즘의 적용을 위해 PI 제어기와 결합된 RL 제어기의 설계기준을 실험 및 이론적으로 연구하여 다음과 같은 결론을 도출하였다.

(1) PI 제어기에 RL 제어 알고리즘을 적용할 경우 냉난방 모두의 경우에 빠른 응답성을 보이고 제어기의 성능을 개선시킬 수 있다.

(2) 겨울철, 여름철 조건에서는 강화신호 비율이 설정치 추종능력에 영향을 주며 강화신호 비율이 낮은 경우 추종능력이 좋아지는 결과를 나타낸다. 따라서 냉난방 시스템에 동시에 적용되는 PI 제어기와 복합된 RL 제어기를 설계할 때에는 강화신호 비율을 낮게 설계해야 한다.

후 기

본 연구는 NRL의 재정적 지원을 받아 수행되었으며, 이에 감사를 드립니다.

참고문헌

1. Ministry of Commerce, Industry and Energy, 2003, Total energy consumption report, pp. 1-80.
2. Virk, G. S. and Loveday, D. L., 1992, A comparison of predictive, PID, and on/off techniques for energy management and control, Proceedings of ASHRAE, pp. 3-10.
3. Hang, C. C. and Åström, K. J. and Ho, W. K., 1991, *Refinements of the Ziegler-Nichols* tuning formula, IEE Proceedings Part D-Control Theory Applicat., Vol. 138, No. 2, pp. 111-118.
4. Watkins, C. and Dayan, P., 1992, Technical note: Q-learning, Machine Learning, Vol. 8, pp. 279-292.
5. Anderson, C. W., Hittle, D. C., Katz, A. D. and Kretchmar, R. M., 1997, Synthesis of reinforcement learning, neural networks, and PI control applied to a simulated heating coil. Artificial Intelligence in Engineering, Vol. 11, No. 4, pp. 421-429.
6. Anderson, C. W., 1993, Q-learning with hidden-unit restarting, Advances in Neural Information Processing Systems, Vol. 5, S. J. Hanson, J. D. Cowan and C. L. Giles, eds., Morgan Kaufmann Publishers, San Mateo, CA, pp. 81-88.
7. Barto, A. G., Bradtke, S. J. and Singh, S. P., 1995, Learning to act using real-time dynamic programming, Artificial Intelligence, Special Volume: Computational Research on Interaction and Agency, Vol. 72, No. 1, pp. 81-138.

8. Sutton, R. S, 1988, Learning to predict by the method of temporal difference, Machine Learning, Vol. 9, pp. 9-44.
9. Sutton, R. S. and Barto, A. G., 1998, Reinforcement Learning: an Introduction, Cambridge, MA: MIT Press, pp. 51-85.
10. So, J. H., Cho, S. H., Song, M. H. and Park, M. S., 2001, Experimental study on control performance of reinforcement learning method, Proceedings of the SAREK, pp. 697-701.