적응형 미들웨어의 자가 진단을 위한 베이지안 네트워크를 사용한 진단엔진

최보윤[O] 김경중 조성배
연세대학교 컴퓨터과학과
bychoi@sclab.yonsei.ac.kr[O], kjkim@cs.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

A Diagnosis Engine Using Bayesian Network for Self-management of Adaptive Middleware

Bo-Yoon Choi[O], Kyung-Joong Kim, Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

분산 어플리케이션은 동시에 여러 사용자가 각기 다른 환경에서 동기화된 프로세서를 사용하기 때문에 일정한 성능을 유지하는 것이 무엇보다 중요하다. 진단엔진은 시스템을 진단하여 시스템 결함의 원인을 발견하여 시스템이 자가치료가 가능하게 한다. 적응형 미들웨어는 진단엔진을 사용해서 분산 어플리케이션이 로컬환경에 맞는 고른 서비스를 유지할 수 있도록 한다. 본 논문은 베이지안 네트워크를 사용한 적응형 미들웨어의 진단엔진을 제안한다. 베이지안 네트워크는 상황인지분야에서 널리 사용되는 추론기법으로서, 수집된 데이터를 통해서 그 구조를 학습하고 데이터를 증거 값으로 시스템 진단을 한다. 본 논문은 실험 대상자로부터 윈도우 시스템에서 두 시간 동안 데이터를 수집하여 한 시간은 베이지안 네트워크 학습에 사용하고, 나머지는 베이지안 네트워크 성능평가에 사용하였다. 실험 결과 학습된 두 개의 베이지안 네트워크 모델은 각각 95.41%, 99.77%의 정확성을 보였다.

1. 서 론

분산 시스템이 네트워크를 통하여 단일화되고 컴퓨팅 환경이 복잡해질수록, 인터넷을 사용한 서비스와 같이 분산환경에서 이루어지는 서비스를 지원하는 어플리케이션의 성능을 고르게 유지하는 것이 중요하다. 이런 대부분의 문제는 어플리케이션이 실행되는 실시간에 일어나는 문제로 빠른 시간 안에 문제를 해결해야 한다. 이러한 문제를 해결하기 위해서는 시스템상황을 파악할 수 있는 진단 엔진이 필요하다. 진단엔진은 시스템 성능을 분석하고, 결점을 파악하며, 시스템에 변화가 필요한가를 결정한다. 미들웨어는 분산환경에서 이용되는 분산어플리케이션의 통신과 데이터 전송의 역할을 하고 있다.

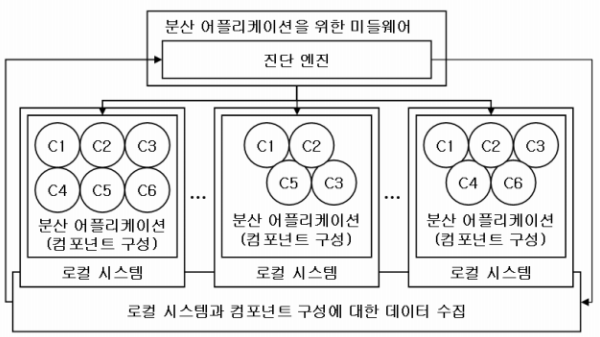


그림 1. 진단엔진을 사용한 적응형 미들웨어

그림 1은 진단엔진을 사용한 적응형 미들웨어의 구조이다. 적응형 미들웨어는 로컬 시스템과 컴포넌트 기반으로 만들어진 분산어플리케이션의 컴포넌트 구성에 대한 정보를 수집하고, 수집된 정보와 진단엔진을 사용하여 각 로컬 환경에 대한 진단을 한다. 미들웨어는 진단결과를 사용하여 분산어플리케이션의 컴포넌트 구성을 결정하여 다양한 로컬환경에 맞는 서비스를 지원한다. 베이지안 네트워크는 상황인지분야에서 널리 사용되는 추론기법으로 적응형 미들웨어에서 진단엔진으로 사용되었다. 이것이 로컬시스템의 진단과 분산어플리케이션의 성능분석을 한다. 본 논문은 적응형 미들웨어를 위한 베이지안 네트워크를 사용한 진단엔진을 제안한다. 진단엔진은 미들웨어가 각 로컬환경의 분산어플리케이션에게 적절한 서비스를 할 수 있도록 할 것이다.

2. 제안하는 진단엔진

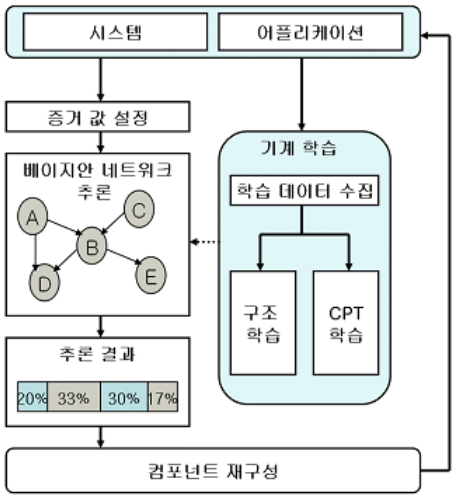


그림 2. 베이지안 네트워크를 사용한 진단엔진의 구조

그림2는 베이지안 네트워크를 사용한 진단엔진이다. 진단엔진

은 분산어플리케이션이 실행되면 로컬환경과 어플리케이션에 대한 정보를 수집한다. 수집된 정보는 전처리 과정을 거쳐 일정 기간 동안 수집된다. 수집된 데이터는 베이지안 네트워크 구조와 확률 값 학습에 사용된다. 학습한 베이지안 네트워크는 로컬 환경과 어플리케이션으로부터 데이터를 증거 값으로 하여 시스템을 진단하며, 그 결과로 로컬환경에 적절한 어플리케이션의 컴포넌트를 재구성하여 안정된 서비스를 제공한다.

2.1. 데이터 수집

시스템 정보와 사용자의 행동패턴은 시스템 진단의 중요한 정보이다. 시스템 정보는 네트워크, 메모리사용량, CPU 사용률, 실행중인 프로세서의 수 등이 될 수 있다. 사용자 행동패턴은 어플리케이션의 실행 유무를 말하며, 시스템의 사용유무나, 사용 시간 등이 시스템진단의 정보로 사용될 수 있다. Eric Horvitz는 컴퓨터의 사용유무를 관찰하여 사용자의 행동패턴을 모델링했다[1].

2.2. 베이지안 네트워크

베이지안 네트워크를 표현하기 위해서는 $<B,\theta_B>$를 사용한다. $<B,\theta_B>$는 B 라는 베이지안 네트워크의 θ_B 라는 확률변수를 말한다. $P<B,\theta_B>$는 베이지안 네트워크의 모든 변수에 대한 결합확률분포(Joint Probability Distribution)를 말한다. 베이지안 네트워크는 방향성 비순환 그래프로서 노드와 그 노드들 간의 의존관계를 나타내는 아크로 이루어졌다. $B=(V,E)$ 는 베이지안 네트워크 B 의 노드 집합 $V=\{x_1,x_2,...,x_n\}$ 와 그들의 의존관계를 나타낸 E 를 말한다. $x_i \in V$ 인 모든 변수에 대해, $P(x_i \mid Pa(x_i))$ 라는 조건부 확률을 가진다. 여기서 $Pa(x_i)$ 는 노드 x_i 의 부모노드 집합을 말한다.

$$P<B,\theta_B>=P(x_1,x_2,...,x_i)=\prod_{i=1}^{n}P(x_i \mid Pa(x_i))$$

베이지안 네트워크를 사용한 확률추론은 불충분한 정보를 가진 상황에서 추론할 때 많이 사용되는 방법이다[2][3]. 추론하기 위해서는, 베이지안 네트워크 구조를 설계하고, 결합 확률분포가 결정되어야 한다. 보통 구조는 전문가에 의해서 설계되거나 학습을 통해서 이루어진다. 조건부 확률은 전문가가 계산하거나 수집된 데이터를 통해서 학습된다. 그림 3에서 x_1,x_2,x_3,x_4,x_5 는 주어진 도메인의 노드이고, $P(x_1,x_2,x_3,x_4,x_5)$ 는 그들의 결합 확률분포인데, 변수들의 독립성 가정과 체인룰을 이용하여 다음과 같이 구할 수 있다[3].

$$P(x_1,x_2,x_3,x_4,x_5)=P(x_1)P(x_2)P(x_3 \mid x_1,x_2)P(x_4 \mid x_2)P(x_5 \mid x_3)$$

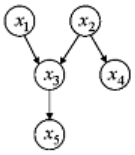


그림 3 베이지안 네트워크 구조

3. 실험 및 결과

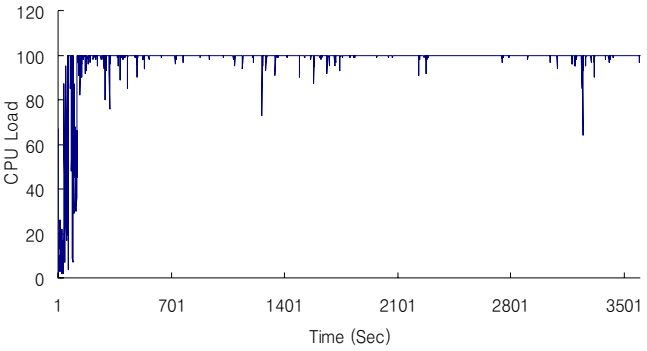
베이지안 네트워크를 학습하기 위해 데이터 수집을 하였다. 데이터수집 대상자는 프로그래머와 일반학생이다. 모니터링 프로그램을 설치한 후, 실험대상자들은 일상적인 업무를 보았다. 실험대상자들이 업무를 보는 동안, 모니터링 프로그램은 실행중인 어플리케이션 리스트와 CPU사용률을 수집하였다. 표 1은 실험대상자들이 두 시간 동안 주로 했던 작업과 실행된 어플리케이션의 개수, 수집된 데이터의 개수이다. 학생 실험대상자는 다른 로컬 시스템의 사용자와 계속적으로 온라인 게임을 했고, 프로그래머 실험대상자는 계속적으로 프로그래밍 업무를 보면서, 데이터베이스에 접근했다. 온라인 게임과 데이터베이스 질의작업은 모두 분산 어플리케이션으로서 적응형 미들웨어를 사용한다면 로컬 시스템 상태에 적절한 데이터 전송을 할 수 있다. 이들은 그 외에도 여러 프로그램을 실행하면서 다른 업무를 보았으며, 실험 대상자가 실행한 프로그램 외에도 시스템이 실행하는 프로세서들이 실행되었다.

표 1. 두 시간 동안 수집된 데이터 정보

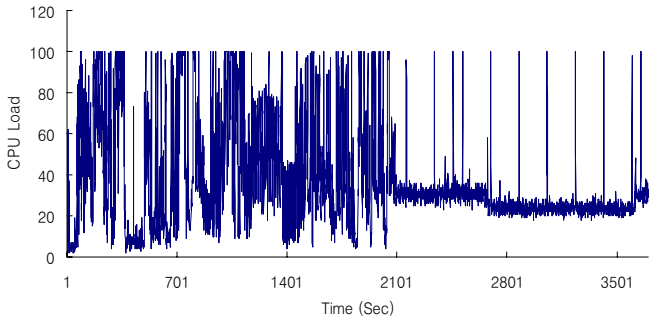
	시간	주요 업무	실행된 프로세서 개수	샘플의 개수
사용자 1	11:39:01 – 14:13:14	온라인 게임	39	7148
사용자 2	11:25:51 – 13:28:43	DB 질의작업, 프로그래밍	41	7194

표 2. 실행된 프로세서 리스트

	프로세서 리스트
사용자 1	SystemIdleProcess, System, smss, csrss, winlogon, services, lsass, ati2evxx0, svchost0, svchost1, svchost2, svchost3, svchost4, spoolsv ,ati2evxx1, explorer, V3monnt, V3monsvc, myLinker, DrVirus, ctfmon, msmsgs, trayapp, alg, LCDPlyer, wscntfy, CDSLicenseMng, conime, taskmgr, CpuUsage, ProcMon, TalesWeaver, InphaseNXD, IEXPLORE, skcbgm, npkagt, regsvr32, npkcsvc, npdownv
사용자 2	SystemIdleProcess, System, smss, csrss, winlogon, services, lsass, svchost, spoolsv, explorer, MsgPlus, monsysnt, v3p3at, daemon, ctfmon, ProtHelp, ClientSM, conime, ahnsdsv, ahnsd, AszTray, monsvcnt, v3impro, msnmsgr, putty, TOAD, notepad, editplus, CpuUsage, ProcMon, EMEDITOR, IEXPLORE, v3syson, rundll32, npcopyv, npdownv, NPMON, logon, sucer, supdate, autoup

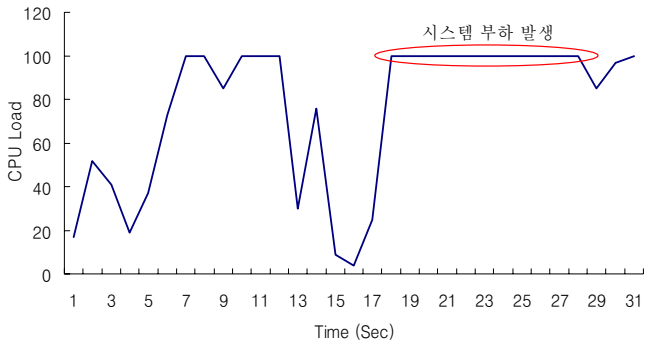


(a) 사용자 1

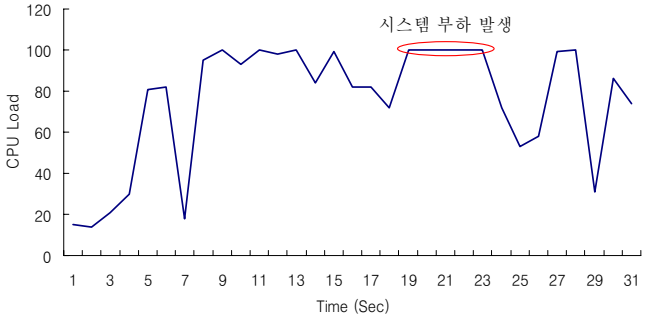


(b) 사용자 2

그림 4 학습데이터의 CPU 사용률

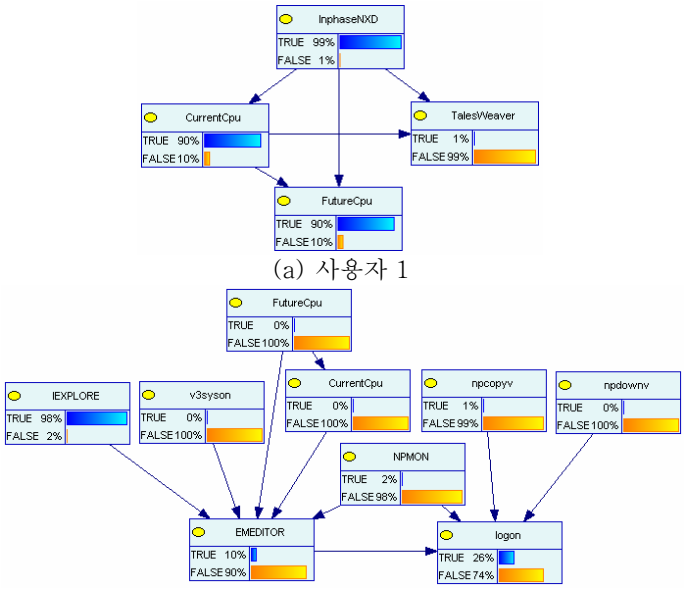


(a) 사용자 1(38초: 2006/1/8 11:39:51-2006/1/8 11:40:29)



(b) 사용자 2(41초: 2006/1/2 11:31:02-2006/1/2 11:31:43)

그림 5 시스템 부하 발생 상황



(a) 사용자 1

(b) 사용자 2

그림 6 학습된 베이지안 네트워크

실험에서 실험 대상자들은 평소 사용하던 개인 컴퓨터를 사용하였다. 데이터는 두 시간 동안 수집되었으며, 한 시간은 학습을 위한 데이터로 나머지 한 시간은 테스트를 위한 데이터로 사용되었다. 실험에서 사용할 베이지안 네트워크는 현재 CPU사용률과 시스템에 실행되고 있는 프로세서 리스트를 사용해서 가까운 미래의 CPU상태를 추론한다. 미래의 CPU상태를 추론하기 위해서 베이지안 네트워크는 현재 CPU정의와 가까운 미래의 CPU상태를 정의해야 한다. 현재 CPU노드와 미래 CPU노드는 CPU사용률이 100%는 5초 이상 유지할 때와 그렇지 않을 때로 나누어 두 가지 상태로 정의하였다. 이는 CPU사용률이 100%를 5초 이하로 유지하는 경우는 웹 브라우저를 실행시킬 때처럼 다른 프로그램을 처음 실행하면 일어나는 일시적인 현상으로 프로세서가 실행되고 나면 사라지는 현상이다. 이런 경우는 시스템 전체에 영향을 주지는 않는다. 반면, CPU사용률이 100%로 장기간(5초 이상)지속될 경우 시스템 성능에 영향을 줄 수 있다. 그림 4는 사용자 1과 사용자 2의 학습데이터의 CPU 사용률을 보여주고 있다. 그림 5는 그림4의 서브 그래프이다. 그림 5는 CPU 사용률이 100%로 5초 이상 지속되어 시스템 성능이 감소되어 분산어플리케이션에 영향을 받을 경우를 보여준다. 그림 6은 데이터를 사용하여 학습한 베이지안 네트워크의 구조이다. 시스템 진단을 위한 베이지안 네트워크는 사용자의 컴퓨터 사용패턴에 의존하기 때문에 사용자 1의 구조와 사용자 2의 구조가 다르다. 표 3은 설계된 베이지안 네트워크의 테스트 데이터에 대한 성능을 보여준다. 사용자 1의 베이지안 네트워크는 95.41%의 정확성을 보여주었고, 사용자 2의 베이지안 네트워크는 99.77%의 정확성을 보여주었다.

표 3 베이지안 네트워크 성능표

	CPU 사용률 100%가 5초 이상	CPU 사용률 100%가 5초 미만	정확성
CPU 사용률100%가 5초 이상	1275	83	93.88 %
CPU 사용률100%가 5초 미만	81	2135	96.34 %
데이터 개수	1356	2218	95.41 %

(a) 사용자 1

	CPU 사용률 100%가 5초 이상	CPU 사용률 100%가 5초 미만	정확성
CPU 사용률 100%가 5초 이상	17	3	85.00 %
CPU 사용률100%가 5초 미만	5	3572	99.86 %
데이터 개수	22	3573	99.77 %

(b) 사용자 2

4. 결론 및 향후 연구

본 논문은 적응형 미들웨어를 위한 베이지안 네트워크를 사용한 진단엔진을 제안하였다. 진단엔진은 사용자의 컴퓨터 사용패턴과 CPU상태를 베이지안 네트워크를 사용하여 확률적 의존관계로 표현하였다. 실험결과 베이지안 네트워크는 사용자 1에 대해 95.41%의 정확성을 보였고, 사용자 2에 대해 99.77%의 정확성을 보였다. 향후에는 다양한 상황을 진단할 수 있도록, 단일 구조의 베이지안 네트워크가 아닌, 다양한 베이지안 네트워크를 사용한 모델의 연구가 필요하다.

참고 문헌

[1] E. Horvitz, C. M. Kadie, T. Paek, and D. Hovel, "Models of attention in computing and communications: From principles to applications," *Communications of the ACM*, vol. 46, no. 3, pp. 52-59, 2003.

[2] E. Charniak, "Bayesian networks without tears," *AI Magazine*, vol. 12, no. 4, pp. 50 -63, 1991.

[3] T. Stephenson, "An introduction to Bayesian network theory and usage," *IDIAP-RR00-03*, Institut Dalle Molle d'Intelligence Artificielle Perceptive, 2000.