Journal of Digital Contents Society Vol. 8 No. 3 Sept. 2007(pp. 309-314)

분산형 데이터마이닝 구현을 위한 의사결정나무 모델 전송 기술

김충곤*, 우정근**, 백성욱***

요 약

분산형 데이터마이닝을 위해 의사결정나무 알고리즘은 분산형 협업 환경에 적합하도록 변환되어야 한다. 본 논문에서 제시된 분산형 데이터마이닝 시스템은 각각의 사이트에서 부분적인 데이터를 위한 데이터마이닝 작업을 수행할 수 있는 에이전트와 여러 에이전트들의 협업을 통해 최종적인 의사결정나무 모델을 완성할 수 있도록 에이전트들 간의 통신을 중재하는 미디에이터로 구성되어 있다. 분산형 데이터마이닝의 장점 중에 하나는 여러 사이트에 분산되어 있는 대량의 데이터를 분산 처리하므로 데이터마이닝의 소요시간을 현저하게 줄일 수 있다는 점이다. 그러나 각 사이트들에 존재하고 있는 에이전트들 간의 통신에 부하가 과도하게 걸린다면, 효율적인 시스템으로의 활용도가 낮아질 것 이다. 본 논문은 에이전트들 간에 의사결정나무 모델의 전송량을 최소로 할 수 있는 방법론에 초점을 맞추었다.

The Transfer Technique among Decision Tree Models for Distributed Data Mining

Choong Gon Kim*, Jung Geun Woo**, Sung Wook Baik***

Abstract

A decision tree algorithm should be modified to be suitable in distributed and collaborative environments for distributed data mining. The distributed data mining system proposed in this paper consists of several agents and a mediator. Each agent deals with a local data mining for data in each local site and communicates with one another to build the global decision tree model. The mediator helps several agents to efficiently communicate among them. One of advantages in distributed data mining is to save much time to analyze huge data with several agents. The paper focuses on a transfer technique among agents dealing with each local decision tree model to reduce huge overhead in communication among them.

Keyword : Distributed Data Mining, Decision Tree, Agent, Mediator

1. 서론

디데이터 마이닝은 데이터웨어하우스나 데이터마트 안에 저장되어 있는 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 대량의 데이터는 분석 없이 그 자체만으로는 의미가 없다. 그 이유는 그 데이터의 용량이 너무 방대하고 복잡해서 보통 사람들이 그 데이터를 이해하기란 매우 힘들기 때문이다. 데이터 분석가들도 그들이 필요한 부분만을 선택하여 심층 분석을 하기도 한다. 그러나 그 이외에 사용하지 않은 데이터들로부터의 의미 있는 정보나 지식 등이 발견 되지 않는 경우가 있기 마련이다. 이러한 지식을 유출해내는 방법은 어떤 특정 기법과 그 기술 자체만을 의미하는 것이 아니고, 비즈니스 문제나 과학용 데이터들을 분석하는 문제 등에서 주어진 상황을 이해하고 그 특정한 문제를 해결하기 위하여 여러 학문 분야의 방법을 적용하는 포괄적인 과정을 의미한다. 인터넷과 분산데이터베이스의 발달로 인한 온라인 데이터들의 급증으로 분산형 데이

※ 제일저자(First Author) : 김충곤
접수일자:2007년08월01일, 심사완료:2007년08월17일
* 세종대학교 컴퓨터공학부
forgom@gmail.com
** 세종대학교 컴퓨터공학부
*** 세종대학교 컴퓨터공학부(교신저자)
▣ 이 논문은 2006년도 신기술 연구개발 지원 사업의 지원에 의하여 이루어진 것임(과제번호-10643)

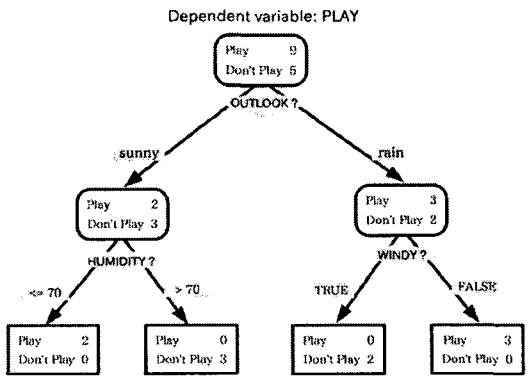
터 마이닝의 필요성은 더욱 절실해졌다. 분산 데이터베이스 시스템의 사용 범위 및 활용도는 계속 증가하고 있는 추세이며 그 이유는 다음과 같다[1].

첫째, 대부분의 조직들은 구조상 분산된 형태를 취하고 있기 때문에 분산 데이터베이스 접근이 분산된 조직에 쉽게 응용될 수 있다. 둘째, 몇 개의 데이터베이스가 이미 분산된 조직 내에 존재하는 경우, 분산데이터베이스 시스템의 사용은 분산된 조직에 보다 자연적인 접근이 된다. 셋째, 현재의 조직이 새로운 조직들을 합병할 경우, 분산 데이터베이스 접근은 현재의 조직에 최소한의 영향을 미치면서 점증적인 성장을 지원할 수 있다. 넷째, 지역적으로 분산 된 조직 내에서 분산 데이터베이스 시스템은 중앙집중식 데이터베이스 시스템 보다 데이터 통신량 면에 있어서 효율적으로 운영될 수 있다. 그러므로 지역적으로 분산된 조직들이 통신망으로 연결되어진 경우 분산 데이터베이스 시스템은 정보를 공유함에 있어서 보다 효율적인 대안이 되고 있다. 분산 데이터베이스 시스템에서는 데이터 파일을 필요로 하는 모든 곳에 똑같은 데이터 파일을 설치할 필요가 없다. 데이터 파일은 통신망에 의해 다른 곳에서도 접근이 가능하기 때문이다. 똑같은 데이터 파일을 여러 곳에 설치할 경우 데이터 파일을 저장하기 위한 비용이 증가 할 뿐만 아니라, 어떤 곳에서 데이터파일 갱신 요구가 있는 경우 일관된 데이터 파일을 유지하기 위하여 모든 곳에 분산된 데이터 파일을 갱신하여야 하는 추가비용이 들어가게 된다. 그러나 데이터 파일을 한 곳에만 보관할 경우 위에서 언급한 비용은 발생하지 않지만 데이터 파일이 없는 곳에서 데이터 파일에 접근(질의)하기 위해서는 통신 비용을 지불하여야 한다. 그러므로 분산 데이터베이스 시스템에서 해결해야 하는 하나의 문제는 전송비용을 최소화하기 위하여 데이터파일의 전송량을 줄여야 하는 점이다[2,3,4,5,6].

분산형 데이터 마이닝은 분산된 지역에 각각의 정보를 가지고 있고 데이터 마이닝 엔진을 가지고 있다. 각각 데이터 마이닝 엔진을 이용해 데이터 마이닝을 하고 필요한 의사결정나무 모델 정보를 다른 지역에 보내거나 받는다. 본 논문은 데이터마이닝에 사용되는 의사결정나무 알고리즘을 분산형 데이터마이닝에 사용하기 위하여 분산 환경에서 의사결정나무 모델정보를 전송하는 방법에 대해 초점을 맞췄다.

2. 분산형 데이터 마이닝

의사결정나무 알고리즘은 일반적인 데이터마이닝 알고리즘중 하나이다. 의사결정나무 알고리즘은 데이터를 구성하는 속성의 수가 불필요하게 많을 경우에도 분류에 영향을 미치지 않는 속성들을 자동으로 제외시키기 때문에 데이터 선정도 용이하다. 그리고 어떠한 속성들이 각각의 결과에 결정적인 영향을 주는지도 쉽게 파악할 수 있다. 다음 (그림 1)은 의사결정나무 알고리즘을 이용하여 날씨와 사람의 패턴을 분석한 사례이다.

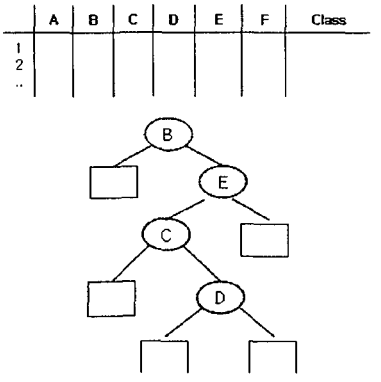


(그림 1) 의사결정나무의 사례

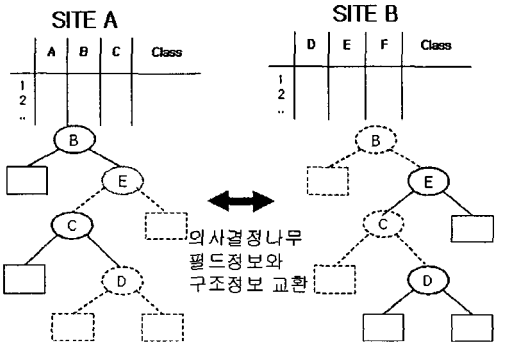
데이터 마이닝에서는 데이터의 양이 많으면 많을수록 유용하면서도 믿을만한 결과가 나올 가능성이 커진다. 분산된 환경에서 데이터 마이닝이 가능하다면 데이터를 더 많이 모을 수 있고 이는 마이닝 결과가 유용한 결과를 낼 수 있다.

현실 세계에서 데이터들은 일반적으로 분산되어 있다. (그림 2)는 중앙 집중 환경에서 의사결정나무 알고리즘을 이용하여 모델을 완성하는 모습이다. 기존 싱글 데이터마이닝을 이용하여 분산되어 있는 데이터를 분석하기 위해서는 분산되어 있는 데이터를 모두 한 장소로 모아야만 했다. 하지만 분산형 데이터마이닝을 이용할 시 분산되어 있는 데이터를 이동할 필요 없이 네트

워크를 이용하여 의사결정나무 모델의 필드정보와 모델정보의 교환만으로 분석이 가능하다. (그림 3)은 분산 환경에서 사용할 수 있도록 변형된 의사결정나무 알고리즘을 이용하여 모델을 완성하는 모습이다.

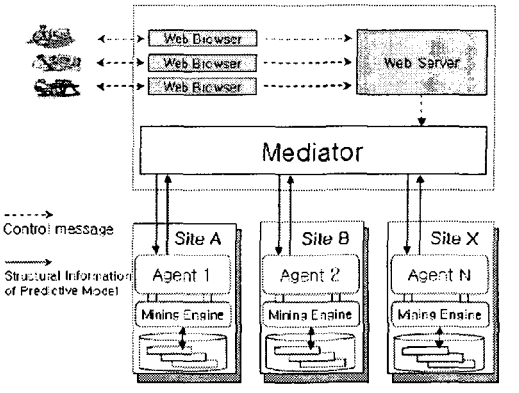


(그림 2) 중앙집중식
의사결정나무 모델의 완성



(그림 3) 분산 환경에서 의사결정나무
모델의 완성

분산된 환경에서 데이터 마이닝을 하기 위해서는 각 분산된 장소를 나타내는 에이전트와 그 에이전트를 관리할 수 있는 미디에이터가 필요하다. 에이전트는 마이닝 엔진을 가지고 있고 미디에이터는 중계적인 역할을 한다. (그림 4)는 미데이터와 각 분산된 사이트 간의 시스템 모델 정보를 나타내고 있다. 각 에이전트는 모두 에이전트 엔진을 가지고 있으며 미디에이터는 이들 에이전트 사이의 정보전송을 도와주는 역할만 한다.



(그림 4) 에이전트 기반의 분산형 데이터
마이닝 시스템 구조

의사결정나무 알고리즘을 이용한 분산형 데이터 마이닝은 다음 9단계에 걸쳐 수행된다[7].

1. [미디에이터] 각 에이전트에 데이터 마이닝 작업을 시작을 알린다.
2. [에이전트] 자신의 사이트에서 마이닝 작업을 하는 동안 데이터를 가장 최적으로 나눌 수 있는 애트리뷰트와 그 기준 값을 값을 찾는다.
3. 단계 2에서 찾은 최적의 정보를 미디에이터로 보낸다.
4. [미디에이터] 각각의 장소에서 보내온 값들을 비교하여 최적의 정보를 선택한다.
5. [미디에이터] 각각의 에이전트에게 트리를 나눌 것인지 아니면 기다릴 것인지에 대한 정보를 보낸다.
6. [에이전트] 미디에이터에서 최종 선택된 값을 가진 에이전트는 데이터를 분리하여 의사결정트리의 일부분을 완성하고 분리정보를 미디에이터로 보낸다.
8. [미디에이터] 분리정보를 최종 선택된 값을 받지 못하여 기다리고 있는 에이전트에게 보낸다.
7. [에이전트] 미디에이터로부터 받은 분리정보를 이용하여 의사결정트리의 일부분을 완성한다.
8. [에이전트] 의사결정트리가 모두 완성이 되어 더 이상 나눌 수 없다면 미디에이터에 종료를 통보하고, 의사결정나무가 완성되지 않

았다면 단계 2 로 간다.

9. [미디에이터 & 에이전트] 데이터 마이닝 작업 종료

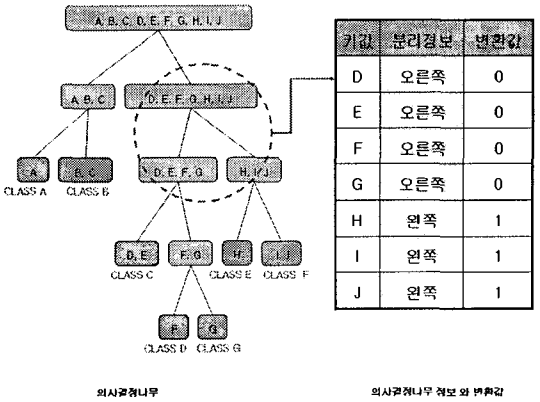
3. 시스템 설계 및 구현

많은 양의 데이터를 마이닝 하기 위해서는 정보를 한곳에 모두 모아야 한다. 만약 모든 데이터를 한곳에 모아도 데이터양이 너무 거대하면 많은 시간과 컴퓨터 자원을 소비하게 된다. 반면 실행시간과 컴퓨터자원을 분산시킨다면 데이터 마이닝을 실행 시에 많은 자원을 절감 할 수 있게 된다. 분산된 환경에서 데이터마이닝을 하기 위해서 각각의 장소에서 데이터의 전송이 효율적으로 이루어져야 한다. 자원을 분산시켰는데 교환해야할 정보의 양이 너무 클 경우 오히려 집중된 마이닝 형태보다 더 비효율적이 된다. 전송량을 최소로 줄이기 위해 전송할 내용을 가장 작은 단위로 표현하고 가장 작은 단위로 표현된 정보를 한 번 더 변환을 한다.

3.1 분산형 데이터 마이닝 전송 방법 구현

위 분산형 데이터마이닝 알고리즘 9단계 중 6단계에 있는 의사결정나무 모델정보의 표현 방법을 구현 하고자 한다. 의사결정나무 알고리즘을 이용한 분산형 데이터 마이닝에서 키 값은 항상 왼쪽 혹은 오른쪽 값만 가진다. 이처럼 두 가지 값 중 하나의 값으로만 나오는 것을 이용하여 왼쪽으로 분리된다는 정보를 가질 때는 0, 오른쪽으로 분리된다는 정보를 가질 때는 1의 값으로 표현하여 전송한다.

(그림 5)의 변환 방법을 이용하면 "오른쪽, 오른쪽, 오른쪽, 오른쪽, 왼쪽, 왼쪽, 왼쪽" 으로 전송해야할 정보를 "0000111" 으로 전송이 가능하다. 결국 의사결정나무 정보로 변환된 정보는 0과 1로 나타나게 되는데 일반적으로 저장 할 경우 char 1byte로 전송한다. 하지만 0과 1은 bit로 표현되므로 char 1byte에 8개의 bit로 저장하여 보낼 수 있다. 이렇게 정보를 bit로 저장하여 전송을 할 경우 전송량이 1/8로 줄어든다.

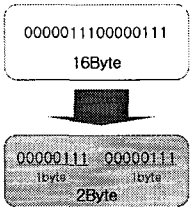


키값	분리경보	변환값
D	오른쪽	0
E	오른쪽	0
F	오른쪽	0
G	오른쪽	0
H	왼쪽	1
I	왼쪽	1
J	왼쪽	1

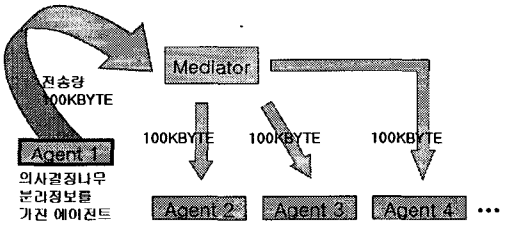
(그림 5) 의사결정나무 모델정보 변환

3.2 의사결정나무 모델정보 전송량과 속도 측정

에이젠트가 가지고 있는데 의사결정나무 모델정보의 크기가 100Kbyte라고 가정했을 경우 실제로 전송되는 전송량의 크기는 이보다 더 커지게 된다. 에이전트에서 미디에이터로 보내고 미디에이터에서 다시 다른 에이전트로 보내다. 처음 전송량은 100kbyte지만 다른 에이전트 간의 전송량까지 고려한다면 Agent가 4개일 경우 400kbyte가 된다. 실제 전송량을 계산해보면 의사결정나무 모델정보가 한번 전송될 때 실제 전송량은 (Agent의수 * 전송량) 이다. 따라서 전송량은 Agent가 늘어남에 따라 계속 늘어나게 된다[7].

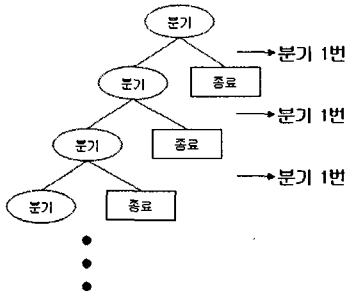


(그림 6) 의사결정나무 모델정보 저장

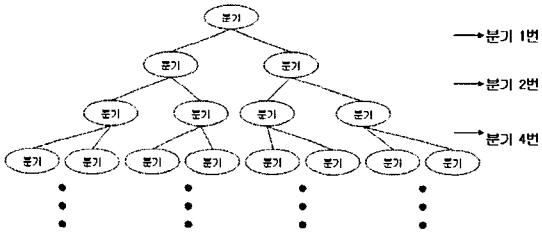


(그림 7) 미디에이터와 에이전트 간 전송량

의사결정나무 정보를 보낼 때는 최적의 경우와 최악의 경우를 생각해볼 수 있다. (그림 8)은 의사결정 알고리즘 최적의 경우를 나타낸다. 최적의 경우 한번 의사결정나무가 분리될 때마다 노드가 그대로이므로 전송횟수가 변함이 없다. 반면 (그림 9)에서처럼 의사결정나무가 분리 될 때 노드수가 2배로 증가하게 되므로 전송횟수가 늘어나게 되어 전송시간이 늘어나게 되므로 최악의 경우가 된다. 분산형 데이터 마이닝에서는 각각의 에이전트에서 최적의 값을 미디에이터로 보내고 미디에이터에서는 다시 한 번 최적의 값을 뽑아내는 알고리즘을 통해 전송량이 최상으로 근접하도록 설계 되어 있다[8].

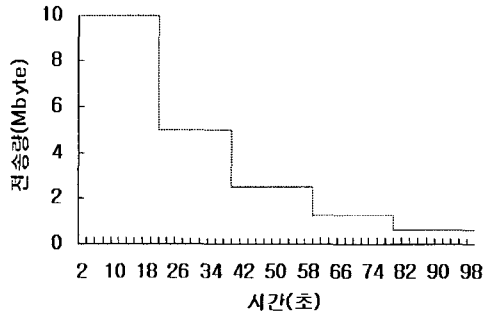


(그림 8) 의사결정나무 알고리즘 최적의 경우



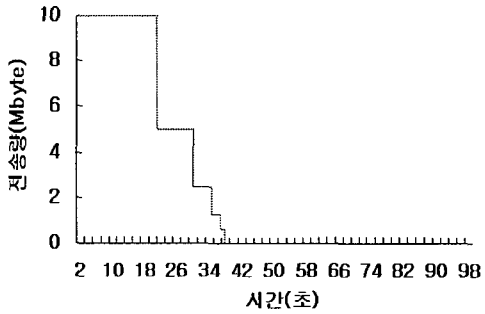
(그림 9) 의사결정나무 알고리즘 최악의 경우

먼저 최악의 경우를 생각해보면 class가 모두 동시에 끝나는 경우다. 중간에 class가 결정되지 않은 경우 한 번에 전송할 전송량은 반으로 줄어들지만 노드가 2배로 늘어나기 때문에 (그림 10)처럼 한 번에 전송되는 양은 증가하지만 시간은 일정하게 된다.



(그림 10) 의사결정나무 모델정보의 시간에 따른 정보 전송량

두 번째로 최적의 경우 노드가 분리될 때마다 한쪽의 class가 결정되는 경우다. 이 경우엔 전송량이 절반으로 줄고 시간도 절반으로 줄게 된다.



(그림 11) 의사결정나무 모델정보의 시간에 따른 정보 전송량

4. 결론 및 향후 연구방향

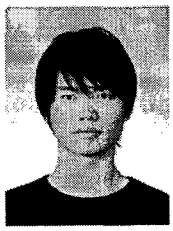
인터넷 환경의 급속한 변화로 많은 양의 정보가 각각의 장소에서 계속 쌓이고 있다. 그리고 해마다 정보를 저장한 사이트들이 급격히 증가하고 있고 분산형 데이터 마이닝으로 얻을 수 있는 정보의 양이 계속해서 늘어나고 있다. 갈수록 정보의 양이 기하급수적으로 증가하는 현대사회에서 분산형 데이터 마이닝을 실행하기 위해서는 분산 환경의 네트워크 속도를 증가시키는 방법과 분산형 데이터 마이닝에서 전송하는

정보를 줄이는 방법이 있다. 첫 번째 방법은 비용도 많이 들뿐더러 현실적으로 힘들기 때문에 전송 정보를 최소한으로 만드는 것이 중요하다. 일반적인 회선에서 800Mbyte정보를 전송 시에 1시간 정도 걸린다고 가정하였을 경우 본 논문에서 제시한 분산형 데이터 마이닝 정보 전송 시스템을 통해 얻은 결과를 이용하면 7.5분(1/8 감소)으로 시간을 줄일 수 있다. 그러나 데이터 전송량이 테라바이트 단위로 커진다면 본 논문에서 제시한 방법으로도 전송하는데 다소 무리가 있다. 이렇게 많은 양의 데이터를 전송하기 위해서 압축을 하는 방법을 다양화하여 정보 전송량을 줄이기 위한 연구가 필요할 것이다.

참 고 문 헌

1. S. Ceri and G. Pelagatti, "Distributed Databases - Principles and Systems", McGraw Hill, 1984
2. 서필교, "분산 데이터베이스시스템에서 전송량 최소화를 위한 데이터파일 배치 및 경영보고서 작성위치 결정 모형", 經濟論叢, Vol.18 No.1, 1999
3. Lin Wujuan and Bharadwaj Veeravalli, "An object replication algorithm for real-time distributed databases", Lecture Notes in Computer Science, vol.19, pp. 125-146, 2006
4. s. Pappe, W. Effelsberg and W. Lamersdorf, "Database Access in Open Systems", Lecture Notes in Computer Science, vol.248, pp.148-164, 2006
5. Mariella Di Giacomo, Mark Martinez, and Jeff Scott, "A Large-Scale Digital Library System to Integrate Heterogeneous Data of Distributed Databases", Lecture Notes in Computer Science, vol.3149, pp.391-397, 2004
6. Takao Mohri and Yuji Takada, "Virtual Integration *of Distributed Database by Multiple Agents*", Lecture Notes in Computer Science, vol.1532, pp.391-397, 1998
7. Sung Wook Baik, Jerzy Bala and Ju Sang Cho, "Agent Based Distributed Data Mining", Lecture Notes in Computer Science, vol.3320, pp. 42-45, 2004
8. Sung Wook Baik, Jerzy Bala and Ju Sang Cho, "Performance Evaluation of an Agent Based Distributed Data Mining System", Lecture Notes in Computer Science, vol.3501, pp. 25 - 32, 2005

김 충 곤

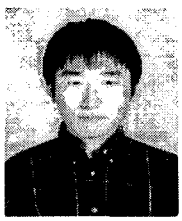


2006년 : 세종대학교 디지털콘텐츠학과(학사)

2006년~현 재 : 세종대학교 디지털콘텐츠학과(석사)

관심분야 : 데이터마이닝, 데이터베이스

우 정 근



2007년~현 재 : 세종대학교 디지털콘텐츠학과(학사)

관심분야 : 데이터마이닝, 데이터베이스

백 성 욱

1987년 : 서울대학교 계산통계학과(학사)

1992년 : Northern Illinois University, Computer Science (석사)

1999년 : George Mason University, Information Technology (박사)

현 재 : 세종대학교 디지털콘텐츠학과 교수

관심분야 : 컴퓨터비전, 데이터마이닝, 데이터베이스, 이미지프로세싱, 컴퓨터 게임, 가상현실, 디지털콘텐츠