이러닝을 위한 클러스터 기반 학습 자원의 저장 기법

Storing Method of Learning Resources based on Cluster for e-Learning

윤홍원
신라대학교 컴퓨정보공학부

Hong-Won Yun(hwyun@silla.ac.kr)

요약

SCORM에서 학습 자원은 공유 가능 콘텐츠 객체 또는 하나 이상의 애셋(asset)으로 구성된다. 이러닝 환경에서 애셋을 신속하게 검색하고 재사용할 수 있는 저장 방법이 필요하지만 아직 관련된 연구가 거의 없다. 본 논문에서는 클러스터에 기반을 둔 애셋의 저장 방법을 제안하고 수학적으로 정형화하여 정의하였다. 또한, 애셋을 평가하는 기준과 각 애셋을 평가하는 절차를 제시하였다. 실험을 통하여 제안한 클러스터 저장 방법에 기반을 둔 검색이 텍스트 카테고리화에 기반한 검색보다 처리시간과 정확도 측면에서 성능이 우수함을 보였다.

■ 중심어 : | 이러닝 | 클러스터 | 저장 기법 |

Abstract

A learning resource is a SCO or a collection of on or more assets in the SCORM. A storage policy is required to search rapidly and reuse assets in e-learning environment. However there are not research results about it. In this paper, We propose a storing method for assets based on cluster and define the mathematical formulation of it. Also, we present criteria for assets evaluation and describe procedure to evaluate each asset. We show that the search based on proposed cluster storing method increase performance than the categorization search through performance evaluation.

■ keyword : | e-Learning | Cluster | Storing Method |

I. 서 론

ADL의 SCORM(Sharable Content Object Reference Model)은 컴퓨터 및 웹 기반 학습의 기술적인 프레임웍 안에서 재사용 가능한 학습 콘텐츠를 생성하는데 목적을 두고 있다. SCORM은 콘텐츠 집합 모델(Content Aggregation Model)과 실행 환경(Run-Time Environment)을 정의하고 학습자의 요구에 따라 다양하게 학습 콘텐츠를 제공하기 위한 시퀀싱과 네비게이션을 정의하고 있다[1].

학습 관리 시스템은 학습 콘텐츠를 학습자에게 전달하고 관리하는 역할을 수행한다. SCORM에서 학습 관리 시스템이 학습자에게 전달하고 추적할 수 있는 학습의 가장 작은 논리적 단위를 공유 가능 콘텐츠 객체 (SCO: Sharable Content Objects)라고 하고, 가장 기본적인 학습 콘텐츠는 웹 클라이언트에게 제공할 수 있는 미디어, 텍스트, 이미지, 사운드, 웹 페이지 등의 전자적인 표현인 애셋으로 구성된다[1-4].

접수번호 : #061215-001
접수일자 : 2006년 12월 15일
심사완료일 : 2007년 01월 10일
교신저자 : 윤홍원, e-mail : hwyun@silla.ac.kr

현재 SCORM은 기본적인 학습 순서 모델을 제시하고 있으나[3] 학습자가 학습 내용을 검색하는 경우에 검색한 학습 내용을 제시하는 모델이나 방법에 대한 연구 결과는 아직 나오지 않고 있다. 학습 자원의 기본 요소인 애셋은 메타 데이터로 설명이 되므로 학습 자원의 사용과 재사용을 위해서 체계적으로 검색될 수 있다. 따라서 애셋의 저장 방법을 개선한다면 학습 내용의 체계적인 검색은 물론 검색 속도를 향상시킬 수 있다.

지금까지 학습 자원의 저장 방법에 대한 연구는 거의 없으나 인접한 연구로써 텍스트 카테고리화를 적용한 클러스터 기법에 관한 연구가 있다[5-7]. 이 연구들은 문서에 나오는 텍스트를 카테고리화하여 저장하고 검색하는데 적합하지만 재사용이 가능한 이러닝 환경에서 학습 자원의 빠른 검색을 위한 저장 방법으로 적합하지 않다.

본 논문에서는 분산 환경에서 재사용이 가능한 학습 자원을 효율적으로 저장하는 방법을 정의한다. 실험을 통하여 제안한 클러스터 저장 방법에 기반을 둔 검색이 텍스트 카테고리화 방법보다 성능이 우수함을 보인다. 본 논문의 구성은 다음과 같다. 2장에서는 본 연구와 관련된 연구를 소개하고 3장에서는 학습 자원의 효율적인 저장 방법을 정의한다. 4장에서는 실험 결과를 분석하고 마지막으로, 5장에서 결론을 맺는다.

II. 관련된 연구와 학습 자원

1. 텍스트 카테고리화

텍스트 카테고리화는 온라인 문서의 급속한 증가로 인하여 텍스트 처리 기술의 핵심이 되었으며 인터넷이나 인트라넷 상의 도큐먼트를 분류하는 데 이용되고 있다. 텍스트 카테고리화는 카테고리 검색의 한 과정으로 볼 수 있고, 그 목표는 주어진 도큐먼트를 이미 정의된 카테고리로 분류하는 것이다. 도큐먼트의 분류의 기본적인 방법은 각 도큐먼트를 기계학습을 통하여 하나 또는 그 이상의 카테고리로 분류한다[5][9][10]. 텍스트 카테고리화와 관련하여 기계 학습에서 도큐먼트의 재표현, 분류 기술, 분류의 평가에 관한 연구가 있다. 또한 도큐먼트 클러스터링에 관한 연구로써 계층적 클러스터링, K-평균 등과 같은 연구 결과가 있었으며, 최근에는 벡터 공간 모델과 단어의 빈도를 이용한 도큐먼트 클러스터링에 관한 연구 결과가 있다.

이러한 연구 결과는 신문기사와 같은 대량의 텍스트를 관리하기에 적절하나 공유 가능한 이러닝의 학습 자원을 저장하는데 이용하기는 적절하지 않다.

2. 학습자원의 저장 단위

SCORM에서 학습 자원은 공유 가능 콘텐츠 객체 또는 하나 이상의 애셋(asset)으로 구성되며 공유 가능한 최소의 학습 단위이다. 각 애셋은 애셋 객체와 애셋을 정의하는 메타 데이터가 있으며 재사용될 수 있다[1][2][8]. 본 논문에서는 학습 콘텐츠를 구성하는 최소 단위인 애셋을 저장 객체로 다룬다. 최근에는 SCORM의 확장에 관한 연구가 진행되고 있으며 이러한 연구에는 콘텐츠 데이터 모델의 구조, 런타임 환경의 설계 등이 포함되어 있으나[1][11] 학습 자원의 저장에 관한 연구 결과는 거의 없다. [그림 1]은 애셋이나 공유 가능 콘텐츠 객체로 구성되는 SCORM의 콘텐츠 구조를 나타낸다.

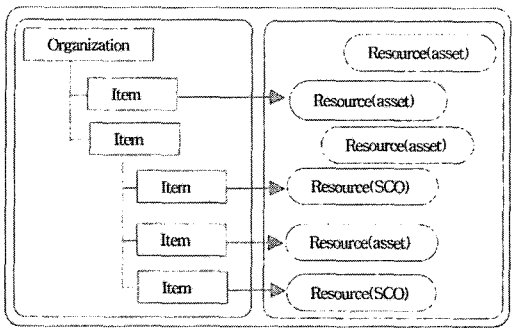


그림 1. SCORM 콘텐츠 구조

III. 이러닝 학습 자원의 저장

1. 학습 자원의 클러스터 정의

본 절에서는 제안하는 이러닝 학습 자원(resource)의 클러스터를 수학적으로 정형화하여 정의한다. $R = \{r_1, r_2, ..., r_N\}$ 는 서로 다른 N개 학습 자원의 집합이라고 한다. 학습 자원 집합의 부분집합 $R_P \subseteq R$로 둔다. 각 학습

자원을 구성하는 애셋은 문자 혹은 비문자 가운데 하나인데, 문자로 표현하는 용어의 집합과 비문자 애셋의 메타 데이터로 사용된 용어의 집합을 포함하는 서로 다른 M개 용어의 집합을 T라 하고 $T = \{t_1, t_2, ..., t_M\}$ 으로 한다. T의 부분집합 $R_T \subseteq T$로 둔다. 하나의 학습 자원을 c로 나타내고 R_T 와 R_P 조합으로 정의한다:

$$c = (R_T, R_P) \quad (1)$$

용어와 학습 자원 공간의 개수가 각각 M, N 이므로 $M \times N$ 행렬 위에 임의의 애셋이 어떤 학습 자원 공간에 있는지 [그림 2]와 같이 나타낼 수 있다. [그림 2]에서 terms와 resource space는 각각 용어와 학습 자원 공간을 나타낸다.

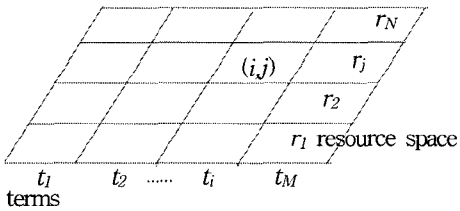


그림 2. 용어와 학습 자원 공간

[그림 2]에서 $M \times N$ 행렬의 (i, j) 번째 셀은 r_j ($\in R$)에 나타내는 t_i ($\in T$)를 뜻한다. t_i가 학습 자원 공간의 집합에 포함된 횟수를 $f(t_i, R)$로 나타낼 수 있고 마찬가지로 r_j에 포함되어 있는 용어의 개수를 $f(T, r_j)$로 나타낼 수 있다. 따라서 $f(t_i, r_j)$는 r_j에 나타나는 t_i의 횟수가 된다. 모든 학습 자원 공간 R에서 모든 T의 횟수를 $f(T, R)$라고 하면 다음과 같다.

$$f(T, R) = \sum_{t_i \in T} f(t_i, R) = \sum_{r_j \in R} f(T, r_j) = \sum_{t_i \in T} \sum_{r_j \in R} f(t_i, r_j) \quad (2)$$

각 용어 t_i ($\in T$)의 확률은 아래와 같다.

$$P(t_i) = \frac{f(t_i, R)}{f(T, R)} \quad (3)$$

사후 확률(posterior probability) $P(r_j|t_i) = P(t_i, r_j) / P(t_i)$ 이므로,

$$P(t_i, r_j) = \frac{f(t_i, r_j)}{f(t_i, R)} \bullet \frac{f(t_i, R)}{f(T, R)} \quad (4)$$

이 사후 확률을 같은 방법으로 R_T와 R_P에 적용하면 다음과 같다.

$$P(R_T, R_P) = \frac{f(R_T, R_P)}{f(R_T, R)} \bullet \frac{f(R_T, R)}{f(T, R)} \quad (5)$$

비문자인 애셋의 메타 데이터로 사용된 용어의 집합을 이루는 서로 다른 L개 용어의 집합을 X라고 하고 $X = \{x_1, x_2, ..., x_L\}$, X의 부분집합 $R_X \subseteq T$로 둔다. 각 비문자인 애셋의 x_i($\in T$)의 확률은 다음과 같다.

$$P(x_i) = \frac{f(x_i, R)}{f(X, R)} \quad (6)$$

비문자인 애셋에 사후 확률을 같은 방법으로 적용하면 다음과 같다.

$$P(x_i, r_j) = \frac{f(x_i, r_j)}{f(x_i, R)} \bullet \frac{f(x_i, R)}{f(X, R)} \quad (7)$$

R_T와 R_P에 같은 방법으로 적용하면 다음과 같다.

$$P(R_X, R_P) = \frac{f(R_X, R_P)}{f(R_X, R)} \bullet \frac{f(R_X, R)}{f(X, R)} \quad (8)$$

2. 학습 자원의 평가 기준

용어와 학습 자원 공간에 대응하는 이산적인 두 개의 임의 변수를 각각 T와 R이라고 하면 확률 이론에서 T와 R사이에 상호 정보는 $I(T;R)$로 표현하고 다음과 같이 계산한다.

$$I(T;R) = \sum_{t_i \in T} \sum_{r_j \in R} P(t_i, r_j) \log \frac{P(t_i, r_j)}{P(t_i) P(r_j)} \quad (9)$$

식(9)에서 $P(t_i)$와 $P(t_i, r_j)$는 각각 식(3)과 식(4)와 같고, $P(r_j) = f(r_j) / f(T, R)$이다. 비문자인 애셋과 학습 자원의 공간에 대응하는 이산적인 두 개의 임의 변수를 각각 X와 R이라고 하고 X와 R사이의 상호 정보는 $I(X;R)$로 나타낸다.

$$I(X;R) = \sum_{x_i \in X} \sum_{r_j \in R} \log \frac{P(x_i, r_j)}{P(x_i) P(r_j)} \quad (10)$$

이 식에서 $P(x_i)$와 $P(x_i, r_j)$는 각각 식(6)과 식(7)과 같다. 비문자인 애셋의 가중치를 ωI로 쓰고 계산하면 $\omega I(T;R) = I(T;R) \times I(X;R)$ 가 된다. 하나의 애셋과 학습 자원의 공간을 조합한 값은 $\upsilon I(T;R)$로 나타내고, $\upsilon I(T;R) = I(T;R) + \omega I(T;R)$로 계산한다. 따라서 $\upsilon I(T;R) = (1 + I(X;R)) \times I(T;R)$ 가 된다. 각 학습 자원의 상호 정보 값은 다음과 같다.

$$I(R_T;R_P) = P(R_T, R_P)\log\frac{P(R_T, R_P)}{P(R_T)P(R_P)} \tag{11}$$

식(11)에서 $P(R_T) = \sum_{t_i \in R_T} P(t_i)$, $P(R_P) = \sum_{r_j \in R_P} P(R_j)$ 이다. 각 학습 자원의 값은 $\eta I(R_T;R_P)$로 표현하고 $\eta I(R_T;R_P) = (1 + I(R_X;R_P) \times I(R_T;R_P))$로 한다. 용어 대 학습 자원 공간의 일대일 대응에서 k개의 임의 변수 S_1, S_2, ..., S_k로 확장하면 다음과 같다.

$$\upsilon I(S_1;S_2;\ldots;S_k) = \sum_{i=1}^{k}(I(S_1;S_2;\ldots;S_k) + \omega I(S_1;S_2;\ldots;S_k)) \tag{12}$$

여러 개 애셋으로 구성된 학습 자원의 평가값 $\upsilon I(S_1;S_1;\ldots;S_k)$를 줄여서 svI 로 쓰기로 한다.

3. 학습 자원 평가의 과정

이러닝 학습 자원을 구성하는 각 애셋을 평가하는 과정은 다음과 같다. 첫째, 각 애셋을 표현하는 용어를 선택하고 그 용어가 들어갈 학습 자원 공간을 선택한다. 각 용어는 문자인 애셋을 대표하는 용어와 비문자인 애셋을 설명하는 메타 데이터를 포함한다. 각 용어와 용어가 포함되는 공유 가능 콘텐츠 객체 공간은 높이가 2인 그래프로 나타낼 수 있으며 어떤 애셋이 어느 공유 가능 콘텐츠 객체 공간에 들어가는지 나타낼 수 있다. 둘째, 각 용어가 대응하는 학습 자원 공간에 나타나는 빈도를 계산한다. 각 용어의 빈도는 $M \times N$의 2차원 행렬로 나타낸다. 앞에서 보인 식(4)로 각 용어의 확률을 계산하고 행렬의 각 셀에 나타낸다. 셋째, 비문자인 애셋을 가지는 용어의 확률을 식(7)을 적용하여 계산한다. 넷째, 용어의 부분집합이 되는 각 학습 자원을 평가한다. 각 학습 자원의 평가는 식(12)를 적용하여 계산한다. 이 식에는 위 셋째 단계에서 구한 비문자인 애셋의 확률을 가중치로 사용한다. 문자로만 구성된 애셋은 가중치가 0이 된다. 다섯째, 평가된 상위 k개의 애셋을 묶어서 클러스터로 한다. 이 학습 자원의 평가 과정은 이러닝 학습 자원을 구성하는 애셋을 클러스터로 묶어서 저장하는 절차와 같다.

IV. 실험 및 결과 분석

1. 실험변수

본 절에서는 텍스트 카테고리화에 기반을 둔 검색과 제안한 클러스터 저장 방법에 기반을 둔 검색 사이의 성능을 비교한다. 본 논문에서 제안한 클러스터 저장 방법은 문자인 애셋을 대표하는 용어와 비문자인 애셋을 나타내는 메타 데이터를 포함하고 있으므로 기존의 근접한 연구인 텍스트 카테고리화와 비교하기가 용이하다. 본 실험에서는 5,000개의 애셋을 대상으로 실험 하였으며 5,000개의 애셋에서 서로 다른 용어는 1,000개가 있다고 가정하였다. 한 개의 애셋에서 중복되는 용어는 최소 1개에서 최대 5개로 하고, 한 개 애셋의 크기는 256KB이고 한 개 용어의 최대 크기는 10 바이트로 하였다. 검색하는 용어를 포함하는 블록을 순차 탐색하였으며 하드디스크의 탐색 시간(seek time)은 3.5ms로 하였다. 애셋의 수를 변화하면서 처리 시간과 정확도를 측정하였고 주어진 시간에서 처리하는 애셋의 수를 비교하였다. 실험을 단순화하기 위하여 처리 시간에 영향이 적은 데이터 전송시간은 성능평가에서 제외하고 성능에 영향이 큰 탐색시간만을 측정하였다.

2. 결과 분석

기존의 텍스트 카테고리화에 기반을 둔 검색 방법은 각 그림의 범례에서 '텍스트카테고리'로 나타내고 제안한 클러스터 저장 기법에 기반을 둔 검색 방법은 '제안기법'으로 표현한다. [그림 3]은 애셋 수의 변화에 따른 처리 시간의 변화를 나타낸다.

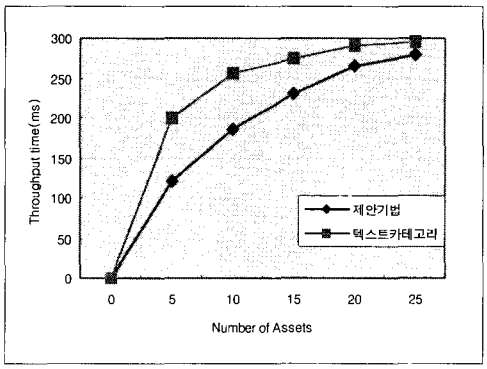


그림 3. 처리 시간

애셋의 수가 5개에서 10개 정도에서 클러스터 저장 방법에 기반을 둔 검색의 성능이 우수하였다. 이것은 확률이 높은 용어를 포함한 애셋이 10개 안팎에서 클러스터링 되어 있기 때문이다. 애셋의 수가 20개 이상이 되면 텍스트 카테고리화 방법과 클러스터 저장 방법에 기반을 검색 사이에 성능 차이가 비슷하게 되었다. 이것은 20개 정도의 애셋을 검색하면 검색 대상인 용어를 포함한 애셋을 거의 다 읽게 되므로 두 방법은 비슷한 성능을 보이게 된다. 학습자가 검색한 용어를 포함한 애셋이 결과로 제출된 애셋의 수 10개 이하에서 발견된다면 제안한 기법은 성능이 우수하다. 학습자는 첫 화면에 제출되면서 상위에 나온 결과를 선택할 가능성이 높으므로 클러스터 저장 방법이 상대적으로 우수함을 알 수 있다.

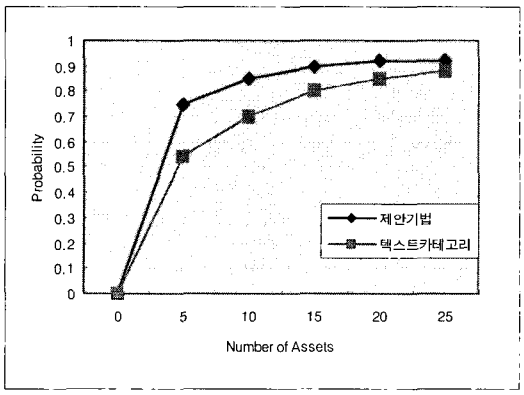


그림 4. 정확도

[그림 4]는 텍스트 카테고리화에 기반을 둔 검색 방법과 제안한 클러스터 저장 방법에 기반을 둔 검색 사이의 정확도를 비교하여 나타낸다. 제안한 방법은 확률이 높고 가중치가 높은 애셋부터 낮은 순으로 결과가 나오고 텍스트 카테고리화 방법은 카테고리 안에서 확률과 상관없이 임의의 것이 선택되어서 결과로 제출된다. 그림 4에서는 검색 결과인 애셋의 수가 적을수록 클러스터에 기반을 둔 저장 방법의 정확도가 높음을 알 수 있다. 제안한 방법은 확률과 가중치가 높은 애셋부터 낮은 순서로 저장되어 있으므로 최초 결과로 제출되는 애셋의 수가 적을수록 정확도가 높고, 텍스트 카테고리와 방법은 선택된 카테고리 안에서 임의로 저장되어 있기 때문에 선택된 애셋의 수가 적을수록 정확도가 낮다.

V. 결 론

본 논문에서는 이러닝 환경에서 재사용이 가능한 학습 콘텐츠를 학습자가 빠르게 검색하기 위한 학습 자원의 저장 방법을 제안하였다. 클러스터에 기반을 둔 학습 자원의 저장 방법을 제안하였고, 수학적으로 정형화하여 정의하였다. 각 학습 자원을 클러스터링하기 위하여 확률과 가중치를 평가하였다. 또한, 학습 자원을 평가하는 기준과 각 학습 자원을 평가하는 절차를 제시하였다. 일반적으로 텍스트 검색에 많이 사용되는 텍스트 카테고리화 방법과 제안하는 클러스터 저장 방법에 기반을 둔 검색 사이에 성능 비교하였다. 실험을 통하여 텍스트 카테고리화 방법과 제안한 방법 사이에 처리 시간, 정확도를 비교하였다. 제안한 클러스터 저장 방법에 기반을 둔 검색이 기존의 검색 방법보다 처리 시간과 정확도의 비교에서 성능이 우수함을 보였다. 이러닝 학습 자원은 확률과 가중치를 이용하여 클러스터로 저장하는 것이, 텍스트 카테고리화 하는 것보다 검색 성능이 우수함을 알 수 있다. 제안한 클러스터 기반 저장 방법은 분산 환경의 이러닝 콘텐츠를 효율적으로 저장하는 방법으로 활용할 수 있으며 학습 콘텐츠의 재사용을 위한 성능 향상의 기법으로 사용할 수 있다.

참 고 문 헌

[1] ADL, SCORM 2004 2nd Edition Overview, p.21, Jul. 2004.

[2] ADL, SCORM Content Aggregation Model, Version 1.3.1 p.11, Jul. 2004.

[3] ADL, SCORM Sequencing and Navigation, Version 1.3.1 p.3, Jul. 2004.

[4] G. Cstagliola, F. Ferrucci, and V. Fuccella, "Web System Architectures: SCORM Run-time Environment as a Service," Proceedings of the 6th International Conference on Web Engineering '06, p.103, Jul. 2006.

[5] M. Iwayama and T. Tokunaga, "Cluster-based Text Categorization: A Comparison of Category Search Strategies," Proceedings of ACM SIGIR'95, pp.273-278, Jul. 1995.

[6] F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, Vol.34, No.1, pp.1-47, Mar. 2003.

[7] R. Bekkerman, R. E. Yaniv, N. Tishby, and Y. Winter, "On Feature Distributional Clustering for Text Categorization," Proceedings of the 24th ACM SIGIR, pp.146-153, Sep. 2001.

[8] ADL, SCORM Run-Time Environment, Version 1.3.1 pp.3-8, Jul. 2004.

[9] X. H. Wang, J. T. Sun, Z. Chen, and C. X. Zhai, "Machine Learning: Latent Semantic Analysis for Multiple-type Interrelated Data Objects," Proceedings of the 29th ACM SIGIR'06, pp.236-243, Aug. 2006.

[10] R. B. Yate and B. R. Neto, Modern Information Retrieval, Addison-Wesley, 1999.

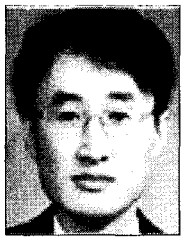
[11] D. Simoes, R. Luis, and N. Horta, "Enhancing the SCORM Metadata Model," Proceedings of the 13th WWW Conference, pp.238-239, May 2004.

[12] Y. J. Li and S. M. Chung, "Text Document Clustering based on Frequent Word Sequence," Proceedings of the 14th ACM CIKM'05, pp.293-294, Oct. 2005.

[13] Y. Zhao and G. Karypis, "Web Clustering: Evaluation of Hierarchical Clustering Algorithms for Document Datasets," Prceedings of the 11th International Conference on Information and Knowledge Management, pp.515-524, Nov. 2005.

저 자 소 개

윤 홍 원(Hong-Won Yun) 정회원



- 1986년 2월 : 부산대학교 계산통계학과 (이학사)
- 1998년 8월 : 부산대학교 전자계산학과 (이학박사)
- 1996년 9월 ~ 현재 : 신라대학교 컴퓨터정보공학부 교수

<관심분야> : 데이터베이스, 시맨틱웹, 이러닝