데이터 마이닝 기법을 이용한 상품 추천 시스템

정민아* · 박경우** · 조성의*

Recommending System of Products based on Data mining Technique

Min-A Jung*· Kyung-Woo Park**· Sung-Eui Cho*

이논문은 2003년도 목포대학교 학술연구지원에 의하여 연구되었음

요 약

전자상거래의 활성화로 인하여 인터넷상에 많은 쇼핑몰이 존재한다. 상품 추천 시스템은 고객이 원하는 정보를 얻기 위해 소요되는 시간과 노력을 절약하기 위해 필요성이 강조되고 있다. 본 논문에서는 고객의 접근 로그 데이터를 분석하기 위해 데이터 마이닝 기법 중 분류 기법을 이용하였다. 접근 로그 데이터는 고객이 쇼핑몰에 접근하였거나 접근하여 상품을 구매한 내역 등에 관한 정보를 포함하고 있다. 제안한 시스템은 두 단계로 구성한다. 제 1단계는 데이터 필터링 모듈과 고객이 접근한 웹 페이지들 사이의 관련성을 추출하는 모듈로 구성하고, 제 2단계는 개인화 모듈과 규칙 생성 모듈로 이루어져 있다. 결과적으로 제안한 시스템은 고객의 패턴을 파악하는데 있어서 고객에게 추천하는 웹 페이지들을 등급화하여 제시함으로써 고객에게 상품 추천을 효율적으로 할 수 있다.

ABSTRACT

There are many e-shopping mall because of revitalization of e-commerce system. It is necessary to recommending system of products that is for saving time and effort of customer. In this paper, we propose the system that is applying classification among data mining techniques to analysis of log data of customer. This log data contains access of user and purchasing of products. The proposed system operates in two phases. The first phase is composed of data filter module and association extraction module among web pages. The second phase is composed of personalization module and rule generation module. Customer can easily know the recommended sites because the proposed system can present rank of the recommended web pages to customer. As a result, the proposed system can efficiently do recommending of products to customer.

키워드

Data mining, Recommending system, Personalization, Data Preprocessing

* 목포대학교 컴퓨터교육과

** 목포대학교 정보공학부 컴퓨터공학전공

접수일자 : 2005. 11. 8

Ⅰ. 서 론

인터넷 쇼핑몰이 활성화되면서 고객은 선택의 폭이 넓어지게 되었지만, 수많은 정보 안에서 고객이 진정으로 원하는 정보를 얻기 위한 시간과 노력이 많이 필요하게 되었다. 따라서 고객이 원하는 정보를 얻기까지 소요되는 시간과 노력을 절약하기 위한 고객 지원 시스템의 필요성이 증가하였다. 고객 지원 시스템의 개발을 위해서 고객의 행동 패턴 분석이 필요하다[1,2]. 개인화는 동적인 웹사이트를 바탕으로 하여 컨텐츠를 제공하며, 또한 차별화된 컨텐츠의 제공을 통하여 고객에 대한 서비스를 극대화한 것이다[3,4]. 개인화 전략이 화두로 등장하면서 이에 대한 해결책으로 추천 시스템이 제시되기 시작하였다[5,6]. 추천 시스템이란 기업이 제공하고자 하는 상품이나 정보 중에서 개별 고객이 선호할 것으로 예상되는 것들을 자동으로 찾아내 고객에게 제시해주는 개인화를 위한 지능적인 소프트웨어 시스템이다. 추천 시스템이 고객의 선호에 맞는 정보를 추천해 주기 위해서는 고객의 성향을 파악해야 한다. 고객 지원을 위해 상품을 추천해 주는 시스템은 고객의 프로파일 정보나 개인화 기술 등을 이용하여 개발되고 있으나 미약한 면이 있다. 고객의 성향을 파악하기 위한 방법으로 웹 로그 파일을 분석하는 연구가 진행되고 있다. 웹 로그에는 고객의 접근 시간, 접근한 웹페이지, 접근 시 사용한 브라우저 등 많은 정보가 포함되어 있다. 많은 정보 가운데서 특정한 규칙을 얻기 위한 마이닝 기법을 적용하기 위해서는 우리에게 필요한 정보만을 추출하고 이 정보를 적용하기 용이한 형태로 변환하는 전처리 작업이 필요하다[7]. 본 논문에서는 전처리 단계에서 마이닝 기법 중 분류 기법을 이용하여 사용자에게 부가적으로 사용자가 방문한 웹페이지 간의 관련성을 제시하기 위해 퍼지 개념을 도입하였다[8]. 전처리 단계를 수행한 후에는 마이닝 기법을 적용하기에 적절한 형태인 트랜잭션들의 모음이 출력된다. 트랜잭션에서 마이닝을 적용하는 트랜잭션의 범위에 따라 생성되는 규칙의 의미 또한 달라지게 된다. 사용자가 로그인하여 로그 아웃하는 세션 기간 동안에 발생하는 규칙을 얻고자 할 때도 있고, 3일 동안에 발생하는 규칙을 얻고자 할 때도 있을 것이다. 그런데, 기존의 시스템에서 트랜잭션 식별 단계에서는 관리자가 원하는 연관 규칙이 있음에도 불구하고 일괄적으로 트랜잭션을 구성하기 때문에 상황에 따른 관리자의 요구사항을 수용할 수 없는 단점이 있다. 본 논문에서는 관리자의 요구사항에 따라 각기 다른 관점의 트랜잭션을 구성함으로써 관리자가 원하는 항목과 범위에 대해서 연관 규칙을 생성할 수 있다. 트랜잭션에 마이닝 기법을 적용하여 유용한 패턴을 생성하였을 때 이 패턴을 기반으로 고객에게 알맞은 상품을 추천해 준다.

본 논문에서는 전처리 단계에서 웹 로그 파일을 일정한 형태로 변환하고, 웹페이지에 대한 관심정도를 등급화해서 제시한다. 또한, 마이닝 기법을 적용하여 얻은 규칙과 고객의 프로파일과 개인화 기술을 이용하여 얻은 규칙을 기반으로 추천 시스템을 구축한다. 제안한 시스템은 웹 로그 파일을 분석한 정보를 첨가하여 추천 시스템의 고객의 행동 패턴 예측을 더욱 향상시킬 수 있다.

Ⅱ. 관련연구

인터넷 쇼핑몰의 경우, 고객들을 유치하기 위한 비즈니스 전략이 필요하며, 고객을 유치시킨 후 계속적인 관계를 유지시킬 수 있는 새로운 개념의 전략이 바로 개인화(Personalization)이다. 개인화를 위한 방법들은 다음과 같다[4].

2.1 공동 필터링(Collaborative Filtering)

동일한 분야에 대해 흥미를 가지는 그룹에 속해 있는 사용자들이 필터링하고자 하는 정보에 대해서 유추된 선호도를 기반으로 정보를 필터링하는 방법을 말한다. 즉 선호도를 비교하여 유사한 취향을 가진 사용자들이 공통적으로 선호할 것이라 예상되는 정보를 추천하는 방식이다. 이를 위해 비슷한 취향을 가진 사용자를 그룹화하는 작업이 필요하다. 하지만 사용자 성향을 학습하기까지 시간이 걸리고, 그룹 안에서 다루지 않은 새로운 상품이 나타났을 경우 추천이 불가능하며 그룹의 사람들과 다른 기호를 지닌 사용자를 수용하지 못한다는 단점이 있다.

2.2 규칙 기반 필터링(Rule-Based Filtering)

인구통계학적 방법 또는 웹에서 사용자들의 답변을 유도하여 얻은 정보를 이용해 만든 규칙을 통해 사용자에게 상품을 추천하는 방식이다. 통제가 용이하고 결과를 쉽게 예측할 수 있으며 시의 적절한 개입을 할 수 있는 반면, 유용한 규칙을 설정하기 위해서는 추천 상품들과 사용자에 대해서 정확하고 깊이 있는 이해가 필요하다. 거짓말을

일삼는 사용자나 한 사용자가 여러 개의 아이디를 갖고 있는 경우, 또는 한 아이디로 여러 사용자가 사용하는 경우에는 정보가 정확하지 않고 그 정보를 통해 만든 규칙도 타당성을 잃게 된다. 한번 정해진 규칙은 지역이나 시대에 따라 자주 변하기 때문에 늘 규칙을 새롭게 만들어야 하는 단점도 있다. 상품과 사용자의 관계를 고려하여 의미 있는 속성을 정의하기 어려운 경우 효과적인 규칙을 설정하는 것도 어려워진다.

2.3 내용 기반 필터링(Content-Based Filtering)

내용 기반 필터링은 추천하고자 하는 상품 또는 정보 자체의 내용과 사용자 프로파일 간의 유사성을 고려하여 추천하는 방식이다. 따라서 사용자의 요구사항이나 성향, 기호 등의 정보를 포함하고 있는 사용자 프로파일을 필요로 한다. 시스템은 사용자가 관심을 보였던 상품이나 정보만을 추천한다. 이러한 방식은 추천 시스템에서 중요한 역할을 해 왔지만, 추천 대상에 대한 내용 정보를 필요로 하고, 사용자 중심적인 추천만이 이루어지기 때문에 추천의 폭이 제한적일 수밖에 없다는 단점이 있다.

2.4 지능형 에이전트(Intelligence Agent)

지능형 에이전트는 인공지능 기술을 이용한 것으로 사용자의 행동에 초점을 맞춘다. 주로 웹에서 사용자의 활동을 관찰하고 사용자가 어떤 내용에 관심을 갖고 있는지를 판단해 정보를 얻는다. 웹에서 이뤄지는 사용자 행동 가운데 특정 페이지를 보는 시간, 인쇄한 페이지, 전자상거래로 구매한 상품 등에 주목한다. 일반적으로 공동 필터링이 효력을 발휘하려면 일정 수 이상의 사용자가 필요한 것에 비해 지능형 에이전트는 사용자가 많지 않아도 적절한 내용을 전달할 수 있다. 분석에 쓰이는 자료가 다른 사용자 자료와 비교를 필요로 하지 않기 때문이다. 그러나 공동 필터링이 사용자가 자신의 선호도에 관한 내용을 입력하는 즉시 사용자에게 맞는 내용을 전달할 수 있는 데 비해 지능형 에이전트는 사용자의 웹 이용 형태를 일정 시간 이상 관찰한 뒤에야 정보 제공이 가능하다는 단점이 있다.

Ⅲ. 데이터 마이닝 기법을 이용한 추천 시스템

본 논문에서 제안한 시스템의 구조는 그림 1과 같다.

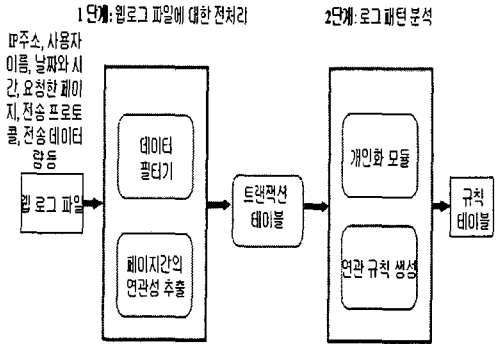

그림 1. 추천 시스템 구조
Fig. 1 The structure of recommending system

먼저, 고객이 웹서버에 접근할 경우, 접근 로그 파일(Access Log File)의 내용은 접근한 사용자의 IP주소, 원격 사용자 로그 인 아이디를 얻기 위한 IdentityCheck 여부, 등록된 사용자 이름, 웹 페이지를 요청한 날짜와 시간, 요청할 때의 메서드, 요청한 웹 페이지, 전송 프로토콜의 종류와 버전, 웹 페이지 전달 성공 여부, 전송 웹 페이지의 데이터 량 등으로 구성된다. 이러한 접근 로그 파일의 내용은 데이터 필터기에 의해 규칙 생성에 필요한 정보만을 추출한다. 추출되는 정보는 등록된 고객 아이디, 접근할 때 고객의 IP주소, 접근한 날짜와 시간, 요청한 페이지를 추출하는 역할을 수행한다. 또한, 페이지 간의 연관성을 추출하기 위해 사용자가 접근한 페이지에 대한 접근 계층도를 퍼지 개념을 적용하여 제공한다. 이러한 페이지 접근 계층도와 필터링된 데이터는 트랜잭션 테이블(Transaction table)에 저장한다.

다음으로 2단계에서는 트랜잭션 테이블의 데이터와 고객의 신상 정보가 저장된 고객 프로파일 테이블(Customer Profile Table)의 데이터와 함께 마이닝 기법 중 분류기법을 적용하여 패턴을 분석한 후에 생성된 규칙을 규칙 테이블(Pattern Table)에 저장한다. 본 논문에서 사용한 분류기법은 결정트리이며 C4.5를 사용하여 규칙을 생성하였다. 또한, 고객이 구매한 상품 정보가 저장된 상품 구매 테이블(Product Purchasing Table) 등의 데이터를 기반으로 개인화 기술을 적용하여 규칙 테이블에 저장한다. 규칙 테이블에 저장된 지식은 메일이나 우편, 사이트에 방문한 고객에게 웹 페이지를 추천해 주는 등 고객을 지원한다.

3.1 웹 로그 파일에 대한 전처리

본 논문에서 제안한 시스템의 1단계에서는 웹 로그 파일에 대한 전처리가 수행된다. 웹서버가 생성하는 로그 파일 중 접근 로그 파일의 형태는 다음과 같다.

61.84.219.150 - ACE [07/oct/2005:17:02:33 +0900] "GET /A.html HTTP/1.1" 200 159

접근 로그 파일의 내용은 접근한 사용자의 IP주소, 원격 사용자 로그 인 아이디를 얻기 위한 IdentityCheck 여부, 등록된 사용자 이름, 웹 페이지를 요청한 날짜와 시간, 요청할 때의 메서드, 요청한 웹 페이지, 전송 프로토콜의 종류와 버전, 웹 페이지 전달 성공 여부, 전송 웹 페이지의 데이터 량, 웹페이지를 통한 구매여부 등으로 구성된다.

표 1. 고객 접근 로그 테이블
Table. 1 Access log table of a customer

CID	Date	Time	Page	Flag
:	:	:	:	
ACE	10/oct/2005	17:02:33	A.html	1
ACE	10/oct/2005	17:17:03	G.html	1
:	:	:	:	
mon	10/oct/2005	17:02:23	A.html	0
mon	10/oct/2005	17:18:42	H.html	1
mon	10/oct/2005	17:19:11	C.html	1
:	:	:	:	

이러한 접근 로그 파일에는 웹 페이지 뿐 아니라 이미지 파일, CGI 파일 등 불필요한 정보까지 기록되어 있다. 데이터 필터기는 표1에서와 같이 시스템의 고객 행동 패턴 분석에 필요한 정보만을 추출한다. 표 1에서 'Flag'는 구매여부를 나타낸다. 페이지 접근 계층도는 고객이 접근한 비슷한 정보에 대한 계층을 나타낸다. 예를 들어 그림 2에서 나타내는 페이지의 계층도는 고객이 관심을 갖는 한가지 정보에 대한 페이지들을 나타낸다. 하위 계층으로 갈수록 검색되는 횟수가 작다는 것을 나타낸다. 이러한 계층도는 고객에게 자세한 정보를 추천하기 위해 부가적으로 제공하게 된다.

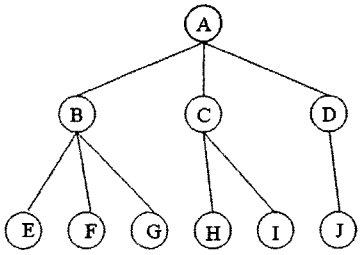

그림 2. 페이지 접근 계층도
Fig. 2 The access hierarchy of a page

페이지 접근 계층도는 추상화 수준을 고려한 결정트리 알고리즘을 적용하여 그림 3과 같이 각 페이지에 따른 소속정도 값을 갖게 된다.

T: 학습 데이터

RID	X	μ(ti)	RID	X	μ(ti)
t_1	a	1.0	t_9	b	1.0
t_2	c	1.0	t_{10}	c	1.0
t_3	d	1.0	t_{11}	f	1.0
t_4	h	0.7	t_{12}	j	0.4
t_5	a	0.9	t_{13}	b	0.6
t_6	c	1.0	t_{14}	d	1.0
t_7	e	0.9	t_{15}	f	1.0
t_8	i	1.0	t_{16}	g	0.9

그림 3. 고객의 접근 페이지 소속정도
Fig. 3 The membership degree of access page

이때, 'RID'는 레코드를 나타내며, 'X'는 고객이 접근한 페이지를 의미하며, $\mu(t_i)$는 유사한 정보 접근을 소속정보를 이용하여 나타낸 것이다. 그림 3에서 결정트리 알고리즘을 위해 임의의 고객에 대한 페이지 접근 정보를 학습 데이터로 간주하였다. 결과적으로 1단계에서는 필터링된 정보와 각 페이지의 소속정도를 포함하는 페이지 접근 계층도가 생성되어 저장된다.

3.2 로그 패턴 분석

제안한 고객별 접근 페이지 소속정도와 필터링된 정보를 이용하여 구축된 결정 트리가 그림 4의 트리라고 가정하자.

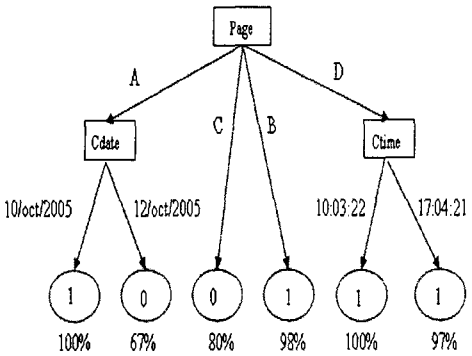

그림 4. 신뢰도를 갖는 결정 트리의 예
Fig. 3 The decision tree having confidence

각 단말 노드에 부착된 레이블은 클래스를 할당하기 위한 신뢰도를 나타낸다[4]. 이는 고객별로 접근한 페이지를 신뢰도에 따라 차례대로 알 수 있으므로 고객들의 구매 성향이나 특정 기간 동안의 구매 성향 등을 분석할 수 있다. 전체 학습 데이터 중 'Page'속성 값이 'A'인 레코드가 85%라고 가정하자. 이러한 분포를 반영하여 뿌리 노드의 첫 번째 가지를 따라 구매여부를 나타내는 'Flag' 속성 값 '1'클래스에 신뢰도 100% × 85% = 85%로 할당될 수 있다. 따라서 '1'에는 85%의 신뢰도로 할당될 수 있으며, A페이지의 접근 계층도의 하위 계층인 B에 대하여 '1'에는 98%의 신뢰도로 할당될 수 있다. 이러한 결정트리를 통한 결과로 생성되는 규칙들을 통하여 고객이 접근하는 유사한 정보에 대한 구매성향을 모두 알 수 있다.

3.3. 개인화 모듈

개인화 기술 방법 중 협동 필터링 모듈은 유사한 취향을 가진 고객들이 공통적으로 선호할 것이라 예상되는 정보를 추천하는 방식이다. 이를 위해 비슷한 취향을 가진 고객들을 군집화하는 작업이 필요하다. 본 시스템은 고객들을 군집화하는 방법으로 고객의 각 속성에 해당하는 다른 고객들의 상품 구매 수를 이용하였다. 고객의 속성은 직업, 연령, 성별을 선택하였다. 먼저 현재 고객과 가장 유사한 구매패턴을 가지는 고객 그룹을 추출하기 위해 고객의 직업코드와 다른 직업코드 간에 가중치를 구한다. 연령코드, 성별코드도 직업코드와 마찬가지로 각각 가중치를 구한다. 앞에서 구했던 직업과 연령에 대한 가중치와 성별의 가중치를 이용하여 세 가중치의 합을 구하게 된다. 협동 필터링 모듈은 세 가중치의 합이 가장 큰 직업과 연령, 성별의 속성을 갖는 고객들로 군집을 형성하고 고객들이 구매했던 상품 데이터들로 트랜잭션을 구성한다. 그리고 트랜잭션 집합을 기반으로 데이터 마이닝을 적용하여 규칙을 생성한다. 이와 같이 협동 필터링 모듈은 고객과 유사한 취향을 가진 고객 그룹의 행동 패턴을 반영하므로 고객의 구매 욕구를 촉진한다.

트랜잭션은 30대 주부인 여성들을 선택하고 상품 소개 페이지를 선택하였다. 이 때 "30대 주부들은 식기세척기를 구입하였을 때 2일 내에 세탁기를 구입한다"와 같은 규칙을 얻을 수 있다. 이와 같은 규칙을 통해서 30대 주부는 경제적으로 여유가 있기 때문에 식기세척기를 구입할 시기에 신혼 때 구매했던 세탁기를 교체한다는 것을 알 수가 있다. 용자에게 맞는 내용을 전달할 수 있는 데 비해 지능형 에이전트는 사용자의 웹 이용 형태를 일정 시간 이상 관찰한 뒤에야 정보 제공이 가능하다는 단점이 있다.

Ⅳ. 결 론

인터넷이 활성화되면서 웹 페이지의 컨텐츠는 증가하고 그에 따라 사용자가 접하는 정보 또한 급격히 증가하게 되었다. 고객은 선택의 폭이 넓어지게 되었지만, 수많은 정보 안에서 고객이 진정으로 원하는 정보를 얻기 위한 시간과 노력이 많이 필요하게 되었고, 고객이 원하는 정보를 얻기까지 소요되는 시간과 노력을 절약하기 위한 고객 지원 시스템의 필요성이 증가하고 있다. 현재 고객을 위한 추천 시스템은 국외에서 뿐 아니라 국내에서도 활발하게 연구되고 웹사이트에 적용하는 사례도 늘고 있다. 이러한 추천 시스템은 고객의 프로파일 정보, 고객의 구매 상품 정보, 개인화 기술 등을 이용하고 있다.

고객이 진정으로 원하는 상품이 무엇인가를 정확히 분석하고자 할 때 다른 기술들의 연구가 더욱 더 필요하다. 본 논문에서는 부가적으로 웹 로그 파일의 데이터를 기반으로 고객이 접근한 웹 페이지들의 관련성을 제시하기 위해 분류 기법을 이용하여 고객의 행동 패턴을 분석하였다. 따라서 고객의 구매 정보와 고객의 프로파일을 기반한 추천 뿐 아니라 고객의 웹 페이지 방문을 기반으로 관심있거나 구매 가능한 상품을 등급화하여 추천할 수 있도록 하였다. 한편, 웹 로그 파일 분석에서 고객들을 식별할 때 로그인 하기 전의 고객들은 식별할 수 없으므로 고객들을 정확히 식별하는 방법에 대한 연구가 필요하다.

참고문헌

[1] M. Bamshad, R. cooley, J. Srivastava, Web Mining : Information and Pattern Discovery on the World Wide Web, *In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*(ICTAI'97), November 1997

[2] M.S, Chen, J. S. Park, P. S. Yu., Data Mining for path traversal patterns in a Web environment, *In proc. 1th International Conference on Distrubuted Computing Systems*, pp385-392, 1996

[3] M. Bamshad, H. Dai, T. Luo, Y. Sung, J. Zhu, Integrating Web Usage and Content Mining for More Effective Personalization, *In Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb2000)*, September 2000

[4] 한정기, 주정애, 윤현준, 개인화란 무엇인가?, *Onbit times*, 2002

[5] D. Kim and S. Kim, Dynamic Expert Group Models for Recommender Systems, *In Proceedings of the First Asia Pacific Conference on Web Intelligence: Research and Development*, Japan, Oct 2001

[6] 이경호, 윤창현, 박두순, 웹마이닝을 이용한 M-commeree 추천 시스템 설계 및 구현, *한국컴퓨터교육학회지*, 제6권 제3호, pp.26-36, 2003

[7] M. Bamshad, R. cooley, J. Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, Vol. 1, No. 1, 1999

[8] Lee, D., Jeong, M., and Won, Y., Decision Trees for Multiple Abstraction Levels of Data, *Lecture Notes in Artificial Intelligence*, 2182, 76-87, 2001

저자소개

정 민 아(Min-A Jung)

1992년 전남대학교 전산통계학과(학사)
1994년 전남대학교 대학원 전산통계학과(이학석사)
2002년 전남대학교 대학원 전산통계학과(이학박사)
2005년 ~ 현재 목포대학교 컴 퓨터 교육과 교수
※관심분야 : 데이터마이닝, 데이터베이스, 생물정보학, 정보보호

박 경 우(Kyung-Woo Park)

1986년 전남대학교 계산통계학과(학사)
1988년 전남대학교 대학원 전산통계학과(석사)
1994년 전남대학교 대학원 전산통계학과(박사)
1995년 ~ 현재 목포대학교 정보 공학부 컴퓨터공학전공 부교수
※관심분야: 분산시스템, 시스템 소프트웨어, 정보보호 etc.

조 성 의(Sung-Eui Cho)

1975년 전남대학교 문리대 수학과(학사)
1981년 전남대학교 교육대학원 수학교육과(교육학석사)
1983년 조선대학교 대학원 전산통계학과(이학석사)
1992년 조선대학교 대학원 전산통계학과(이학박사)
1985년 ~현재 목포대학교 컴퓨 터 교육과 교수
※관심분야: 데이터마이닝, 정보보호, 수치해석