# 신경망 기법을 이용한 온라인 서점 이용자들의 고객 유형 분석

## Analyzing Online Bookstore Customers Using Artificial Neural Network


전현치*, 신영근*, 박상성*, 김명훈**, 장동식*

고려대학교 정보경영공학부*, 건국대학교 산업공학과**

Hyun-Chi Jeon(chichi@korea.ac.kr)*, Young-Geun Shin(toctop@korea.ac.kr)*,
Sang-Sung Park(hanyul@korea.ac.kr)*, Myoung-Hoon Kim(mhkim0114@naver.com)**,
Dong-Sik Jang(jang@korea.ac.kr)*



### 요약

인터넷 기술의 발달로 B2C 전자상거래의 거래액이 꾸준히 증가하고 있으며 이에 따라 인터넷 소비자들의 효과적인 고객관계 관리를 위해서 기업들은 많은 노력을 기울이고 있다. 특히 특정 소비자 집단의 특성을 파악하고 분석하는 것은 효과적인 CRM과 마케팅 전략을 위해서 필수적이다. 따라서 본 논문에서는 온라인 서점을 이용하는 소비자들을 세분화시켜 의미 있는 그룹들로 정의할 수 있는 방법을 제시한다. 설문지를 통하여 데이터를 수집한 후 요인분석을 실시하여 다섯 가지의 주요인을 추출한다. 요인분석 후에 얻어지는 각 응답자별 요인점수를 input 데이터로 하여 군집분석을 실시하고 분류된 6개의 군집과 다섯 가지 주요인들과의 분산분석을 통하여 군집간의 차이성을 검증한다.



■ 중심어 : | 고객관계관리 | 요인분석 | 군집분석 | 신경망 |



### Abstract

Due to the development of internet technology and the steady increase of turnover at B2C market many companies put a lot of work into maintaining a good relationship with internet customers. Particularly, analyzing and understanding specific customer groups are essential for effective CRM and marketing strategy. Thus, this paper proposes the method to define the customers of online bookstore into several meaningful groups. Five important factors and factor scores for each respondent are obtained by Factor Analysis. Six groups are classified by Cluster Analysis and Analysis of Variance(ANOVA) is used to verify the difference between each group.



■ keyword : | CRM | Factor Analysis | Cluster Analysis | Artificial Neural Network |



* 본 연구는 2007년도 두뇌한국 21 사업에 의하여 지원되었습니다.

* 본 연구는 정보통신부 및 정보통신 연구진흥원의 대학 IT 연구센터 지원사업의 연구결과로 수행되었습니다. (IITA-2006-(C1090-0603-0025))

* 본 연구는 정보통신부 및 정보통신 연구진흥원의 IT 신 성장 동력 핵심기술개발사업의 일환으로 수행하였습니다. [2007-S019-01. 정보투명성 보장형 디지털 포렌식 시스템 개발]

접수번호 : #070827-002

접수일자 : 2007년 08월 27일

심사완료일 : 2007년 09월 21일



교신저자 : 장동식, e-mail : jang@korea.ac.kr

# 1. 서 론

## 1.1 연구배경

인터넷 사용 인구의 지속적인 증가로 인하여 산업구조가 빠르게 변하고 있으며 기업들은 인터넷을 통하여 거래하는 고객들과의 관계를 형성하고 유지하기 위해서 많은 노력을 기울이고 있다[1]. 특히 인터넷 소비자들의 기호가 다양해지고 개성화 되면서 특정 소비자집단이 필요로 하는 것을 이해하고 분석하는 것은 효과적인 고객관계관리(Customer Relationship Management)와 차별화된 마케팅 전략(Marketing Strategy)을 위해서 상당히 중요하다[2]. 통계청 자료에 따르면, 2006년 연간 사이버쇼핑몰 거래 규모는 총 13조 4,596억원으로 전년도 10조 6,756억원에 비하여 2조 7,840억원(26.1%)이 증가한 것으로 집계되었다. 이 중 B2C 거래액은 9조 1,315억원으로 전년도에 비해 15.3% 증가하였다[3]. [그림 1]은 연도별 사이버쇼핑몰과 B2C거래액을 보여주고 있으며 인터넷 소비자들의 거래가 꾸준히 증가하고 있는 것을 알 수 있다. [표 1]은 2006년도 기업-소비자간 온라인 몰의 주요 상품 군 거래액을 나타낸 것으로 S/W(게임S/W등) -1254억원(-12.7%), 음반/비디오/악기 -1407억원(-14.8%) 등이 크게 감소하였고, 의류/패션 및 관련 상품 7,886억원(49.8%), 여행 및 예약서비스 4,147억원(25.9%), 아동/유아용품 2,459억 원(60.5%), 컴퓨터 및 주변기기 2,336억 원(22.7%), 서적 1,320억 원(26.6%) 등이 크게 증가하였다.

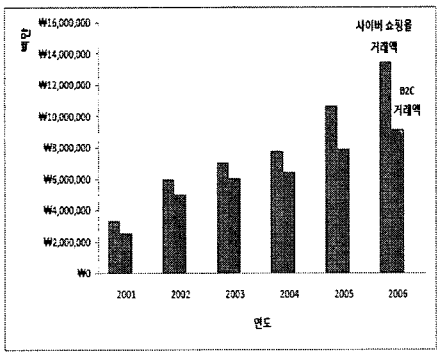



**그림 1. 연도별 거래액 및 증감률 (출처: 통계청)**

**표 1. 2005,2006년 온라인 몰의 주요 상품 군 거래액**

| | 2005년 | | 2006년 | | 전년 | |
|---|---|---|---|---|---|---|
| | | | | | 차 | 증감률 |
| 의류/패션 및 관련상품 | 1,583,101 | 14.8 | 2,371,658 | 17.6 | 788,557 | 49.8 |
| 여행 및 예약서비스 | 1,603,830 | 15.0 | 2,018,509 | 15.0 | 414,679 | 25.9 |
| 아동/유아용품 | 406,365 | 3.8 | 652,264 | 4.8 | 245,899 | 60.5 |
| 컴퓨터 및 주변기기 | 1,026,997 | 9.6 | 1,260,605 | 9.4 | 233,608 | 22.7 |
| 서 적 | 495,666 | 4.6 | 627,675 | 4.7 | 132,009 | 26.6 |
| 스포츠/레저용품 | 395,651 | 3.7 | 501,794 | 3.7 | 106,143 | 26.8 |
| 음반/비디오/악기 | 94,752 | 0.9 | 80,682 | 0.6 | -14,070 | -14.8 |
| S / W ( 게 임 S/W등) | 98,472 | 0.9 | 85,934 | 0.6 | -12,538 | -12.7 |

(단위: 백만 원, %)

B2C 거래에 있어서 도서는 다음과 같은 점에 있어서 향후 인터넷 거래액의 지속적인 증가 가능성을 보여준다. 예를 들어 도서는 제품의 품질이 표준화되어 있고 직접 눈으로 확인하거나 만져 보지 않아도 품질과 내용을 신뢰할 수 있으며 소형이고 특성상 파손될 가능성이 적기 때문에 소비자가 안심하고 구입할 수 있기 때문이다[4]. 따라서 본 논문에서는 온라인 서점을 이용하는 고객들의 효과적인 고객관계 관리를 위하여 소비자들을 비슷한 특성을 가진 집단으로 분류하고 각 집단들의 특성에 따라 고객 유형을 정의하고 세분화 시킬 수 있는 방법을 제시한다.

## 1.2 선행연구

CRM 구축을 위한 고객 유형 분석 및 고객 세분화에 관련된 기존 연구들은 다양한 분야에서 진행되어 왔다. [표 2]는 최근 연구 동향에 대한 내용이다.

**표 2. 최근 연구 동향**

| 저자 | 논문제목 | 기법 |
|---|---|---|
| 조영빈,김채복(2006) | 온라인 소매상점에서의 효과적인 고객 분류 방법론 | 의사결정나무 기법 |
| 홍태호,전성용(2006) | 데이터마이닝을 이용한 고객이탈등급에 기반한 고객 세분화 | 신경망기법(SVM) |
| 이준혁(2006) | 관광유람선 고객의 시장세분화에 관한 연구 | K-means |
| 정현욱,강혜영,김선남(2005) | 인터넷 서점 이용자의 주관성에 관한 Q분석 | Q 방법론 |
| 한진수,서광민(2005) | 호텔 인터넷 이용고객 세분화 | 인구통계학적 특성 |

| | | |
|---|---|---|
| 이영호,강운구,황희정 (2005) | RFM 모델 기반의 병원고객 세분화 전략 | RFM모델 |
| 이정환,최문기(2003) | 고객세분화를 통한 인터넷 쇼핑몰 구매 경험자의 재구매의도 영향 요인 | K-means |

위에서 살펴본 기존 연구들은 주로 통계학적 방법들과 데이터 마이닝 기법들을 이용하여 고객들의 유형을 분석하였다. 본 논문에서는 군집결과를 시각적으로 잘 나타낼 수 있는 자율 신경망 기법인 SOM(Self-Organizing Map)을 이용한 군집분석을 실시하여 클러스터링 기능 향상을 고려한 고객 유형 분석을 실시하였다. 또한 실험에서는 일반적으로 쓰이고 있는 K-means 클러스터링 기법과 비교하였다.

### 1.3 연구방법

본 논문에서는 고객들의 유형을 정의하기 위해서 AIO(Activities, Interests, and Opinions) 접근방법을 이용하여 설문지를 작성하였으며 요인분석을 통하여 설문지들의 타당성과 신뢰성을 검증하였고 이용자들의 특성과 관련된 다섯 가지 주요인을 추출하였다. 요인분석 후에 얻어지는 요인점수를 input data로 하여 군집분석을 실시하였으며 실험을 통해 비슷한 특성을 가진 소비자들을 세분화시키고 세분화된 그룹들과 추출된 다섯 가지 주요인들을 비교하여 각 그룹들에게 영향을 미치는 주요 요인들을 분석하였다. 마지막으로 세분화된 그룹들의 특성을 비교하여 효과적인 고객관계 관리를 할 수 있도록 하였다. 연구 절차는 [그림 2]와 같다.

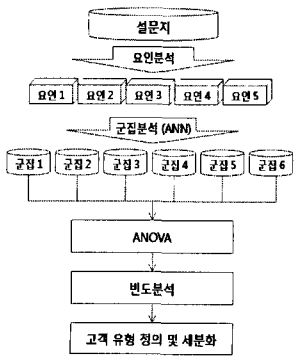



**그림 2. 연구 절차**

## 2. 데이터 수집

2007년 5월 한 달 동안 서울 지역의 대학생 400명을 대상으로 설문지를 수집하였고 설문지의 신뢰성을 높이기 위해서 먼저 pilot survey를 실시하여 이해하기 어려운 질문들을 수정하거나 새로운 항목을 추가하였다. 불성실하게 작성된 14부를 제외한 나머지 386부는 MATLAB 버전7.3.0의 SOM toolbox와 SPSS 12.0을 이용하여 분석하였다. 측정항목들에 대한 척도는 5점 리커트(Likert-scale) 사용하였으며 설문지에 사용된 측정질문들은 Joseph T. Plummer가 제안한 측정항목들을 참고하여 작성되었다[5].

## 3. 데이터 분석

### 3.1 요인분석

요인분석을 실시하기 위해선 먼저 표본자료가 적합한지를 판단하여야 하는데 본 논문에서는 Kaiser-Meyer-Olkin(KMO)측도와 Bartlett의 구형성 검증을 이용하여 표본자료를 검증하였다[표 3].

**표 3. KMO 측도와 Bartlett의 구형성 검증**

| | | |
|---|---|---|
| 표준형성 적절성의 Kaiser-Meyer-Olkin 측도 | | 0.756 |
| Bartlett의 구형성 검정 | 근사 카이제곱 | 1055.148 |
| | 자유도 | 190 |
| | 유의 확률 | 0.000 |

KMO는 표본 적합도를 나타내는 값으로 0.5이상이면 표본자료는 요인분석에 적합하다고 판단할 수 있으며 Bartlett의 구형성 검정은 변수간의 상관행렬이 단위행렬인지 여부를 판단하는 것으로 "모상관행렬은 단위행렬이다"라는 귀무가설을 기각할 수 있어야 변수들 간의 상관관계가 통계적으로 유의하다고 볼 수 있다[6]. [표 3]에서 유의 확률이 0.000이므로 변수간 행렬이 단위행렬이라는 귀무가설이 기각된다. 또한 측정항목들의 신뢰성을 검증하기 위해서 내적일관성 신뢰도(Internal

Consistency Reliability)를 검증하는 크론바하 알파계수(Cronbach's Alpha)를 이용하였다. Cronbach's Alpha 값은 0~1까지의 값을 가지며 α값이 0.7이상이면 높은 신뢰성를 가지는 것으로 볼 수 있고 0.35이하일 경우 해당 항목을 제거하여야 한다[7]. [표 4]는 각 측정 항목들의 신뢰계수 값을 나타내는 것으로 항목이 삭제된 경우 Cronbach's α 값은 해당 항목을 삭제할 경우의 변하게 되는 전체 Cronbach's α 값을 의미한다.

**표 4. 측정항목의 신뢰계수**

| 측정항목 | 항목이 삭제된 경우 Cronbach's α | 전체 Cronbach's α |
|---|---|---|
| Q1 | 0.793 | 0.807 |
| Q2 | 0.792 | |
| Q3 | 0.791 | |
| Q4 | 0.788 | |
| Q5 | 0.793 | |
| Q6 | 0.795 | |
| Q7 | 0.806 | |
| Q8 | 0.791 | |
| Q9 | 0.793 | |
| Q10 | 0.790 | |
| Q11 | 0.800 | |
| Q12 | 0.804 | |
| Q13 | 0.808 | |
| Q14 | 0.798 | |
| Q15 | 0.808 | |
| Q16 | 0.815 | |
| Q17 | 0.796 | |
| Q18 | 0.802 | |
| Q19 | 0.806 | |
| Q20 | 0.802 | |

총 20개의 측정 항목들을 토대로 주성분 분석(Principle Component Analysis)과 베리멕스(Varimax) 요인회전을 통해 요인 분석을 실시한 결과 고유값(Eigenvalue) 1이상을 기준으로 여섯 가지 요인이 추출되었다. 고유값은 요인적재 값의 제곱의 합을 나타내며 고유값이 크다는 것은 그 요인이 변수들의 분산을 잘 설명한다는 것을 의미한다. [그림 3]은 각 요인의 고유값을 나타내주는 스크리 도표(Scree Table)로 각 요인의 설명력이 처음 몇 개 요인까지는 큰 폭으로 감소하다가 어느 위치부터는 감소폭이 매우 체감하는 경향을 보여주며, 감소폭이 체감하기 직전까지의 요인의 수를 기준으로 요인을 추출할 수 있다.

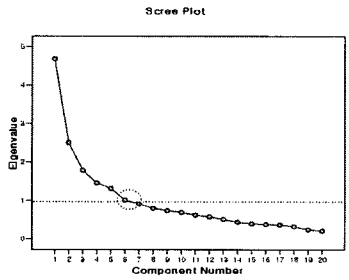



**그림 3. Scree Table**

각 요인에 대한 고유값 및 분산 값은 식(1)과 같이 얻을 수 있다.

$$E = \sum_{i=1}^{Q} (X)^2$$

$$V = \frac{E}{Q} \qquad (1)$$

추출된 여섯 가지 요인들의 고유값 $E$는 주성분 분석을 실시하여 얻어지는 요인행렬 값을 이용하여 얻어진다 [그림 4]. $X$는 20개의 문항 Q대한 각 요인행렬 값을 나타낸다. 분산값 $V$는 고유값 $E$를 전체 문항의 수 $Q$로 나누어서얻어진다.

**Component Matrix[a]**

| | Component | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Q1 | .615 | .012 | -.149 | .006 | -.245 | -.414 |
| Q2 | .606 | .036 | -.177 | -.207 | -.231 | .351 |
| Q3 | .636 | -.199 | -.050 | -.331 | -.007 | .405 |
| Q4 | .718 | -.285 | -.064 | -.368 | -.005 | .173 |
| Q5 | .611 | -.148 | -.182 | -.309 | .227 | .100 |
| Q6 | .563 | -.136 | .098 | .361 | .341 | .054 |
| Q7 | .217 | .604 | -.269 | .337 | .212 | .090 |
| Q8 | .658 | -.167 | -.145 | -.125 | .220 | -.357 |
| Q9 | .555 | .065 | .227 | -.003 | .326 | .001 |
| Q10 | .663 | .039 | -.169 | -.030 | .040 | -.465 |
| Q11 | .458 | -.323 | .223 | .517 | .241 | .046 |
| Q12 | .379 | -.368 | .452 | .309 | -.105 | .007 |
| Q13 | .288 | -.306 | .494 | .270 | -.397 | .181 |
| Q14 | .371 | .391 | .511 | .085 | -.039 | -.031 |
| Q15 | .121 | .612 | .356 | -.105 | .264 | .058 |
| Q16 | .003 | .489 | .496 | -.413 | .227 | -.106 |
| Q17 | .503 | .207 | .020 | .007 | -.616 | -.210 |
| Q18 | .262 | .618 | .297 | -.086 | -.304 | .054 |
| Q19 | .242 | .543 | -.423 | .306 | -.110 | .125 |
| Q20 | .371 | .355 | -.382 | .298 | .021 | .264 |

Extraction Method: Principal Component Analysis.
a. 6 components extracted.

**그림 4. 각 문항에 대한 요인행렬 값**

[표 5]는 식(1)에 의해서 계산된 여섯 가지 요인에 대

한 고유값, 분산 그리고 누적비율을 나타내고 있다.

**표 5. 여섯 가지 요인에 대한 고유값, 분산 및 누적비율**

| 요인 | 고유값 | 분산% | 누적% |
|---|---|---|---|
| 1 | 4.687 | 23.437 | 23.437 |
| 2 | 2.496 | 12.479 | 35.915 |
| 3 | 1.787 | 8.937 | 44.852 |
| 4 | 1.459 | 7.296 | 52.148 |
| 5 | 1.316 | 6.578 | 58.726 |
| 6 | 1.012 | 5.058 | 63.784 |

Varimax 요인회전은 요인행렬의 열(column)의 분산의 합계를 최대화함으로써 요인의 해석을 단순화하는 방법으로 요인 축을 회전함으로써 어떤 변수가 어떤 요인에서 높게 나타는지 알 수 있게 해준다. [그림 5]는 회전된 요인행렬 값을 나타내주고 있으며 각 요인들에 대한 주요 변수들을 확인할 수 있다.

Rotated Component Matrix[a]

| | Component | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Q1 | .167 | .087 | **.857** | -.039 | .123 | .369 |
| Q2 | **.686** | .026 | .083 | -.005 | .229 | .296 |
| Q3 | **.828** | .152 | .075 | .020 | .002 | .077 |
| Q4 | **.787** | .157 | .313 | -.004 | -.084 | .091 |
| Q5 | **.652** | .088 | .356 | .027 | .049 | -.151 |
| Q6 | .209 | **.645** | .258 | .055 | .215 | -.150 |
| Q7 | -.068 | .016 | .090 | .220 | **.768** | -.096 |
| Q8 | .337 | .193 | **.717** | .004 | .007 | -.092 |
| Q9 | .301 | **.383** | .296 | .362 | .079 | -.114 |
| Q10 | .191 | .120 | **.774** | .057 | .153 | .114 |
| Q11 | .077 | **.804** | .151 | -.075 | .076 | -.069 |
| Q12 | .077 | **.684** | .066 | -.092 | -.232 | .261 |
| Q13 | .116 | **.541** | -.152 | -.052 | -.200 | .500 |
| Q14 | .008 | .294 | .086 | **.611** | .090 | .295 |
| Q15 | -.014 | -.010 | -.040 | **.745** | .196 | -.070 |
| Q16 | -.015 | -.179 | .026 | **.807** | -.172 | -.072 |
| Q17 | .134 | -.005 | .355 | .078 | .148 | **.740** |
| Q18 | .078 | -.083 | -.047 | **.584** | .243 | .482 |
| Q19 | .014 | -.097 | .066 | .024 | **.781** | .176 |
| Q20 | .174 | .077 | .086 | -.020 | **.702** | .058 |

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.

**그림 5. 각 문항에 대한 회전된 요인행렬 값**

[표 6]은 최종적으로 추출된 여섯 가지 요인들의 세부내용과 요인 적재량 그리고 공통성을 나타내주고 있다. 요인적재량(Factor Loading)은 각 측정항목들과 각 요인과의 상관관계 정도를 나타내는 값으로 보통 ±0.3 이상이면 유의하다고 보고 ±0.5 이상인 경우는 매우 높은 유의성을 가진 것으로 판단한다. 공통성(Communality)은 추출된 요인들에 의해서 각 변수가 얼마나 설명되는지를 나타내는 것으로 값이 낮은 변수는 요인분석에서 제외시키는 것이 좋고, communality가 0.4이하면 낮다고 판정한다[8]. 누적설명력은 요인 1이 23.437%로 가장 높은 설명력을 가지는 요인으로 나타났다. 또한 요인 2는 12.487%, 요인 3은 8.937%, 요인4는 7.296%, 요인 5는 6.578%, 요인 6은 5.058%를 설명하고 있으며 전체 누적비율은 63.784%로 나타났다. 고객들을 유형별로 정의하기 위해서 요인 1을 가격의존 형, 요인 2는 정보비교 및 서비스추구 형, 요인 3은 정보검색 형, 요인 4는 정보의존 형, 요인 5는 계획구매 형으로 각각 명명하였고 요인 6은 고객 유형을 정의하는데 큰 의미가 없는 것으로 판단되어 군집분석에서 제외되었다.

**표 6. 주성분 분석과 varimax 요인회전을 이용한 요인추출**

| 요인 및 측정 항목 | 공통성 | 요인 적재량 |
|---|---|---|
| **(요인1) 가격의존 형** | | |
| Q2. 나는 여러 온라인 서점을 둘러보면서 가격을 비교한다. | 0.619 | 0.686 |
| Q3. 나는 온라인 서점을 이용할 때 적립금 및 할인쿠폰 사항을 확인한다. | 0.720 | 0.828 |
| Q4. 나는 돈을 절약하기 위해서 온라인 서점을 이용한다. | 0.755 | 0.787 |
| Q5. 온라인 서점은 이용하기가 편리하다 | 0.586 | 0.652 |
| **(요인2) 정보비교 및 서비스추구 형** | | |
| Q6. 온라인 서점은 이미지정보, 동영상과 같은 디지털 컨텐츠 서비스를 더 제공해야 한다. | 0.594 | 0.645 |
| Q9. 나는 도서를 구입하기 전에 다른 독자들의 평가를 확인한다. | 0.470 | 0.383 |
| Q11. 온라인 서점에서는 책에 대한 정보를 더 많이 제공해야 한다. | 0.691 | 0.804 |
| Q12. 온라인 서점은 가격할인 서비스를 더 늘려야 한다. | 0.600 | 0.684 |
| Q13. 온라인 서점의 배송시간이 더 빨라져야 한다. | 0.624 | 0.541 |
| **(요인3) 정보검색 형** | | |
| Q1. 나는 주로 온라인 서점에서 책의 정보를 찾아본다. | 0.633 | 0.667 |
| Q8. 나는 시간을 절약하기 위해서 온라인 서점을 이용한다. | 0.673 | 0.717 |
| Q10. 나는 온라인 서점을 정기적으로 방문한다. | 0.689 | 0.774 |
| **(요인4) 정보의존 형** | | |
| Q14. 나는 월간/연간 도서 판매순위에 관심이 있다. | 0.562 | 0.611 |
| Q15. 나는 친구들이나 주위 사람들이 추천하는 책을 구입하는 편이다. | 0.600 | 0.745 |
| Q16. 나는 주로 베스트셀러 도서를 구입하는 편이다. | 0.719 | 0.807 |
| Q18. 나는 주로 유명 인사나 작가의 책에 관심이 많다. | 0.648 | 0.584 |
| **(요인5) 계획구매 형** | | |
| Q7. 나는 평소 사람들에게 책을 추천하는 편이다. | 0.660 | 0.768 |
| Q19. 나는 평소 사람들과 책에 대해서 대화하는 것을 좋아한다. | 0.653 | 0.781 |
| Q20. 나는 구입하고 싶은 도서품목을 미리 작성해 둔다. | 0.540 | 0.702 |
| **(요인6)** | | |
| Q17. 온라인 서점은 많은 책들을 비교할 수 있어서 좋다. | 0.720 | 0.740 |

### 3.2 군집분석

군집분석(Cluster Analysis)이란 N개의 개체들을 대상으로 관측 변수 값을 기준으로 N개의 개체들 사이의 유사성(Similarity) 또는 비유사성(Dissimilarity)의 정도를 측정하여 개체들을 가까운 순서대로 군집화 하는 통계적 분석방법으로, 주어진 많은 수의 관측개체를 몇 개의 군집으로 세분화함으로써 대상 집단을 이해하고 군집을 효율적으로 활용하고자 하는 것에 목적을 두고 있다[9]. 본 논문에서는 요인분석 후에 얻어지는 응답자 별 요인점수를 input data로 하여 자율 신경망 기법 중에 하나인 자기조직화지도(Self-Organizing Map)를 이용하여 고객들을 세분화 하였다. 또한 실험에서는 군집분석에서 일반적으로 많이 사용되는 K-means 방법과 비교하였다. [표 7]은 다섯 가지 주요인에 대한 응답자별 요인점수이다.

**표 7. 응답자별 요인점수**

| 요인 / (N=386) | 요인1 | 요인2 | 요인3 | 요인4 | 요인5 |
|---|---|---|---|---|---|
| 1 | 1.95442 | 0.77243 | -0.51306 | 1.15635 | 0.90035 |
| 2 | 0.61537 | -1.59526 | 0.60232 | 0.42904 | -1.45791 |
| 3 | 0.91345 | 0.77729 | -1.54131 | 0.59290 | -1.86895 |
| 4 | 1.63814 | -0.26914 | -0.39615 | -1.00332 | 0.35608 |
| 5 | 0.01450 | -1.74077 | 0.18390 | -0.17096 | -0.89382 |
| 6 | -1.51273 | -0.71736 | 1.47711 | 1.91027 | 0.68587 |
| 7 | 1.17742 | 1.77034 | -1.41479 | -1.68020 | 1.68293 |
| 8 | -0.16383 | 0.67035 | 0.46140 | 0.06079 | 0.10495 |
| 9 | 0.11575 | -1.07900 | 2.16643 | 0.67898 | 0.44391 |
| 10 | -0.81308 | 1.15620 | 2.74754 | -0.56392 | -2.65900 |
| ... | ... | ... | ... | ... | ... |
| 386 | -1.10266 | 0.65611 | -1.28412 | -2.19033 | -1.68231 |

#### 3.2.1 K-means

K-means는 잘 알려진 간단한 자율학습(Unsupervised Learning)알고리즘 중의 하나로 미리 정해놓은 클러스터 수를 기준으로 input 데이터와의 거리를 계산하여 값이 최소가 되는 군집에 위치시키는 방법이다. 소속군집과 input 데이터와의 거리 값은 식(2)에 의해서 얻어진다[10].

$$E = \sum_{k=1}^{c} \sum_{x \in Clust\ k} \| X - C_k \|^2 \qquad (2)$$

여기서 $c$는 미리 정해진 군집의 수이며 $C_k$는 군집의 중심 값을 나타낸다. $X$는 군집 $k$에 속한 input 데이터를 의미한다. 결국 input 데이터와 소속된 군집과의 거리 값이 최소가 되는 $E$를 얻는 것이며 총 군집들의 중심 값에 변화가 없을 때까지 반복된다. K-means 알고리즘은 다음과 같은 절차로 군집을 형성한다.

단계 1. 클러스터의 수 $c$를 input으로 하여 6개의 초기 중심 값(seed points)들이 랜덤하게 설정된다. 이때 $C_k$는 초기 군집 중심 값을 나타낸다.

단계 2. 각 개체와 군집중심과의 거리를 계산하여 가장 가까운 그룹에 위치시킨다.

단계 3. 각각의 개체들이 최소 거리의 그룹에 할당된 후 군집중심 값이 다시 계산된다.

단계 4. 군집중심 값에 변화가 없을 때까지 단계 2와 단계 3을 반복한다.

위와 같이 K-means 방법은 임의의 군집의 수 $c$에 따라서 실험 결과 값이 달라지기 때문에 본 논문에서는 최적의 $c$값을 찾기 위해서 Davies-Bouldin(DB) index를 사용하였다[11]. Davies-Bouldin index는 각 클러스터들에 대한 유사성을 나타내주는 값으로 수치 가 낮을수록 좋은 군집형태로 나누어 졌음을 의미한다. 각 클러스터들의 유사성 측정값($R_{ij}$)은 아래와 같은 조건을 만족하여야 한다. $s_i$는 클러스터의 분산을 나타내며 $d_{ij}$는 클러스터의 비유사성을 의미한다.

- $R_{ij} \geq 0$
- $R_{ij} = R_{ji}$
- if $s_i = 0$ and $s_j = 0$ *then* $R_{ij} = 0$
- if $s_j > s_k$ and $d_{ij} = d_{ik}$ *then* $R_{ij} > R_{ik}$
- if $s_j = s_k$ and $d_{ij} < d_{ik}$ *then* $R_{ij} > R_{ik}$

일반적으로 $R_{ij}$는 식(3)과 같이 정의된다. $v_i$와 $v_j$는 각 $i, j$번째 클러스터의 중심 값을 나타내고, $\| c_i \|$ 는 $i$번째 클러스터에 존재하는 요인수를 나타낸다.

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}$$

$$d_{ij} = d(v_i, v_j),\ s_i = \frac{1}{\| c_i \|} \sum_{x \in c_i} d(x_i, v_i) \quad (3)$$

결국 DB index값은 식(4)에 의해서 얻어진다. $n_c$는 총 클러스터의 개수를 의미한다.

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i,\ where$$

$$R_i = \max(R_{ij}),\ i = 1 \ldots n_c \quad (4)$$

[그림 6]에서 보는바와 같이 클러스터 수의 범위를 2~10개로 지정하였을 때 군집 수 6에서 index 값(1.4062)이 최소가 되므로 $C$ = 6으로 군집분석을 실시하였다.

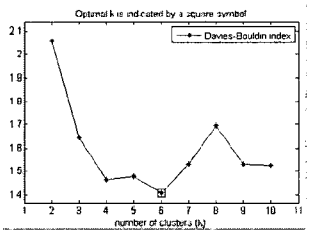



그림 6. DB Index

### 3.2.2 SOM(Self-Organizing Map)

자기조직화지도(Self-Organizing Map : SOM)는 자율 신경망 기법 중의 하나로 Kohonen에 의하여 제안되었다[12]. SOM은 많은 input 데이터들의 관계를 쉽게 파악할 수 있는 시각적인 형태들의 결과를 제공하기 때문에 데이터 마이닝, 음성, 영상 등의 다양한 분야에서 응용되고 있다. 이 알고리즘은 다차원 특징 벡터들을 유사한 패턴끼리 2차원의 특징 지도를 조직화하여 영역 지도를 형성한다. [그림 7]에서 보는바와 같이 다른 비 자율 신경망 기법들과는 달리 SOM은 입력 층과 경쟁 층 2개의 네트워크 구조로 되어있으며 경쟁 층은 2차원의 격자 형태로 구성된다[13].

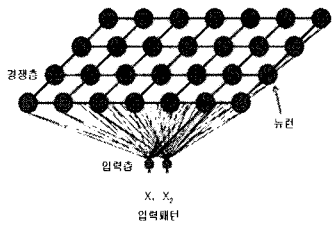



그림 7. SOM 네트워크 구조

입력 층에 입력 벡터 X1, X2가 주어졌을 때 경쟁 층에 있는 각 뉴런들은 prototype vector(표본 벡터)를 의미하고 2차원 격자 공간에 할당된 각 input 벡터 X1, X2에 가장 적합한 가중치 BMU(Best Matching Unit)가 할당된다. [그림 8]은 입력 벡터 x와 표본벡터들의 거리를 계산하여 BMU가 설정되는 그림을 보여주고 있다. BMU는 식(5)와 같이 계산된다. $x$는 입력 벡터, $m_i$는 이웃 표본 벡터들 그리고 $m_c$는 입력 벡터 $x$와 가장 가까운 BMU를 의미한다.

$$\| x - m_c \| = \min \| x - m_i \| \quad (5)$$

BMU가 얻어진 후에 입력 공간의 가중치 벡터들은 BMU가 입력 벡터에 가까워질 수 있도록 갱신되어진다.

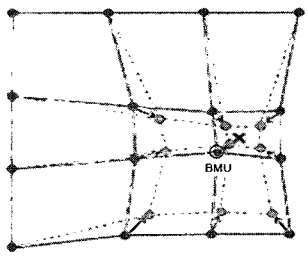



그림 8. Input 데이터 x에 대한 BMU

이러한 새로운 가중치 벡터들의 계산을 위해서 본 논문에서는 SOM의 batch training algorithm을 사용하였으며 식(6)에 의해서 새로운 가중치 벡터들이 얻어진다[14].

$$m_i(t+1) = \frac{\sum_{j=1}^{n} h_{ic(j)}(t) X_j}{\sum_{j=1}^{n} h_{ic(j)}(t)} \quad (6)$$

$c_{(j)}$는 입력벡터 $X_j$에 대한 BMU 값이며 $h_{ic(j)}$는 neighborhood function의 가중치 값 그리고 $n$은 입력 벡터들의 총 수를 의미한다. 여기서 사용된 neighborhood function은 gaussian 함수이며 식(7)과 같다.

$$\exp(-\frac{\| r_c - r_i \|^2}{2\sigma^2(t)}) \quad (7)$$

$r_c$ 와 $r_i$는 격자형태의 map 상에 있는 best matching unit $c$와 뉴런 $i$의 위치를 나타내고 $\sigma(t)$는 $t$시점에서의 표준편차를 나타낸다. 본 논문에서 사용된 SOM 알고리즘 절차는 [그림 9]와 같다.

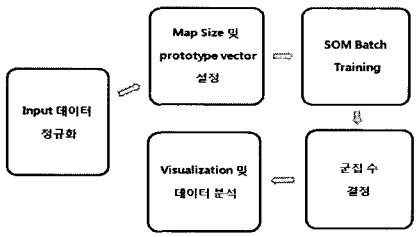



**그림 9. SOM을 이용한 데이터 클러스터링**

위에서 설명한 다섯 가지 응답자별 요인점수들을 SOM의 입력 벡터로 변환하기 위하여 variance normalization을 실시하였다. 데이터들의 평균과 표준편차 값을 구한 뒤 식(8)에 의해서 정규화 된 데이터 값들을 얻을 수 있다. $x_i$는 각 요인점수들을 나타내며 $u_i$는 데이터들의 평균, $\sigma_i$는 표준편차를 의미한다.

$$N_i = \frac{x_i - u_i}{\sigma_i} \quad (8)$$

SOM 알고리즘에 있어서 map unit의 수는 결과 값의 정확도, topographic error, 그리고 quantization error에 큰 영향을 미치므로 데이터 집합에 알맞은 map unit 수의 결정이 필요하다. topographic error는 설정된 BMU가 실제 격자 map내에 근접해 있지 않을 경우를 의미하고, quantization error는 실제 군집 중심과 표본 벡터들과의 차이를 말한다. map size를 크게 하면 할수록 quantization error 값은 낮아지지만 topographic error 값은 올라가는 trade-off 관계를 가진다. 따라서 적절한 map size와 prototype vector는 식(9)에 의해서 얻을 수 있다[15]. $M$은 map unit의 수이며, $N$은 input 데이터의 수이다.

$$M = 5\sqrt{N} \quad (9)$$

본 논문에서 사용된 총 input data는 386이므로 map unit의 수는 약 98개로 map size는 2차원 격자형태의 11x9로 설정 되었다. [그림 10]은 다섯 가지 요인점수에 대한 U-matrix와 각 요인점수별 군집형태를 보여주고 있다. U-matrix는 전체 prototype vector들의 거리를 나타내는 것으로 유사한 색상들은 같은 군집에 할당되는 것을 의미한다. 나머지 변수 5개의 그림은 각 요인점수에 대한 클러스터들의 분포를 나타내는 것으로 각 요인들에 대한 군집 분포들이 다르게 나타나는 것을 알 수 있다. 또한 실험결과 quantization error 값은 1.024 그리고 topographic error 값은 0.067로 나타났다. [그림 11]은 최종 6개의 군집으로 할당된 데이터 군집을 나타내고 있다.

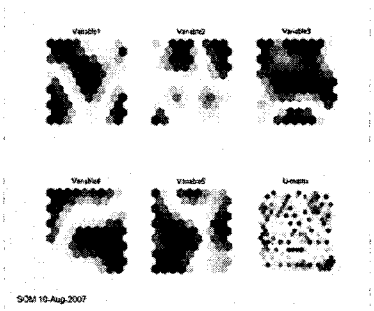


**그림 10. 요인점수별 군집형태 및 U-matrix**

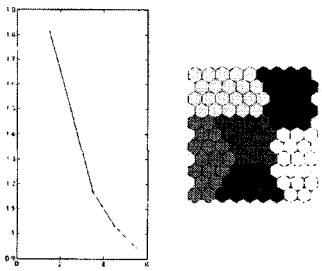


그림 11. 6개의 군집설정

## 4. 실험 결과

SOM과 K-means를 이용하여 온라인 서점을 이용하는 소비자들을 의미 있는 군집들로 분류하였다. [그림 12]에서 보는 바와 같이 군간 평균거리(Inter-Cluster Distance)값이 크면 클수록 그리고 군내 평균거리(Intra-Cluster Distance)값이 작으면 작을수록 좋은 군집형태가 되는 것을 알 수 있다[16].

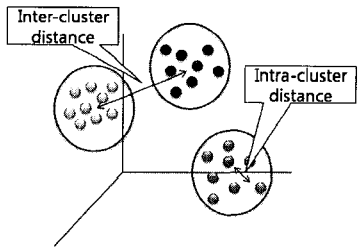



그림 12. 클러스터링 비교

[표 8]과 [표 9]는 각각 SOM과 K-means 방법으로 분류된 여섯 가지 군집들에 대한 평균 inter-cluster와 intra-cluster distance 값을 나타내고 있다. 거리 측정방법은 Euclidean distance를 사용하였다. 대각선 행은 intra-cluster distance 값이며 나머지는 행들은 서로 다른 군집들의 inter-cluster distance값을 의미한다. SOM을 이용하여 군집들의 군내, 군간 평균 거리를 측정한 결과 K-means의 경우보다 전반적으로 intra-cluster distance 값은 낮게 intra-cluster distance 값은 높게 나타났다.

표 8. 평균 군간, 군내 거리 (SOM)

| 군집ID | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1.98 | 3.94 | 3.67 | 3.22 | 3.17 | 3.41 |
| 2 | 3.94 | 2.32 | 3.71 | 3.33 | 3.66 | 3.33 |
| 3 | 3.67 | 3.71 | 2.88 | 3.21 | 3.44 | 3.59 |
| 4 | 3.22 | 3.33 | 3.21 | 2.20 | 3.22 | 2.83 |
| 5 | 3.17 | 3.66 | 3.44 | 3.22 | 2.17 | 2.75 |
| 6 | 3.41 | 3.33 | 3.59 | 2.83 | 2.75 | 2.53 |

표 9. 평균 군간, 군내 거리 (K-means)

| 군집ID | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 2.55 | 2.26 | 2.45 | 2.40 | 2.25 | 2.30 |
| 2 | 2.26 | 3.20 | 3.29 | 3.09 | 3.12 | 3.20 |
| 3 | 2.45 | 3.29 | 2.74 | 3.16 | 3.28 | 3.17 |
| 4 | 2.40 | 3.09 | 3.16 | 3.16 | 2.99 | 2.29 |
| 5 | 2.25 | 3.12 | 3.28 | 2.99 | 2.91 | 2.42 |
| 6 | 2.30 | 3.20 | 3.17 | 2.29 | 2.42 | 3.41 |

따라서 본 논문에서는 SOM을 이용한 군집분석이 더 효과적인 것으로 판단된다. SOM 기법을 이용하여 최종적으로 분류된 여섯 가지 그룹들에 대한 차이성 검증을 위하여 분산분석을 실시하였다. [표 10]은 다섯 가지 주요인과 6개의 군집들에 대한 분산분석을 실시한 결과이다. 6개의 군집 모두 다섯 가지 요인들에 대해 매우 차이가 있는 것으로 검증되었다.

표 10. 분산분석 결과 (SOM, *** P<0.01)

| C / 요인 | 군집 1 54명 | 군집 2 104명 | 군집 3 69명 | 군집 4 28명 | 군집 5 42명 | 군집 6 89명 | F | Sig. |
|---|---|---|---|---|---|---|---|---|
| 요인 1 | -0.76963 | 0.58717 | -0.90154 | -0.75794 | 1.15232 | 0.14306 | 123.89484 | .000*** |
| 요인 2 | -0.11526 | 0.06646 | -1.00820 | 1.02751 | -0.85326 | 0.89183 | 128.86848 | .000*** |
| 요인 3 | -0.65918 | 0.99113 | -0.06618 | 1.07683 | -0.58212 | -0.71108 | 147.21952 | .000*** |
| 요인 4 | -1.33756 | -0.12524 | 0.59007 | -0.09381 | 0.05796 | 0.48912 | 69.60024 | .000*** |
| 요인 5 | -0.39933 | -0.22150 | 0.29951 | 1.09912 | -0.72885 | 0.30167 | 31.54500 | .000*** |

[표10]의 결과를 살펴보면 군집 1은 모든 요인들에 대하여 부정적인 영향을 나타내고 있으며, 군집 2는 요인 3에 대해서 군집 3은 요인4에 대해서 가장 큰 영향을 받

는 것으로 나타났다. 또한 군집 4는 요인2, 요인3, 요인5 세 가지 요인에 대해서 큰 영향을 받으며, 군집 5는 요인 1에 대해서 군집 6은 요인 2에 대해서 큰 영향을 받는 것으로 나타났다.

## 5. 결론 및 향후 연구과제

SOM을 이용하여 분류한 6개의 군집과 다섯 가지 주요인들과의 분산분석 결과를 통하여 온라인 서점 이용자들의 고객 유형을 아래와 같이 정의 할 수 있다.

군집 1: 군집 1에 속한 이용자들은 평소 온라인 서점을 통하여 도서 구입을 잘 하지 않거나 온라인 서점에 관심이 없는 것으로 판단된다.

군집 2(정보검색 형): 군집 2에 속한 이용자들은 주로 책의 정보를 찾아보고, 시간을 절약하기 위해서 온라인 서점을 이용하는 집단이다.

군집 3(정보의존 형): 군집 3에 속한 이용자들은 유명인사들의 광고, 베스트셀러, 판매순위 그리고 주위 사람들의 추천에 의해서 주로 도서를 구입하는 집단이다.

군집 4, 군집 6: 군집 4는 정보비교 및 서비스, 정보검색 형, 계획구매 형 3가지 유형에 속한 집단으로 정보비교, 가격할인 그리고 다른 독자들의 평가를 중요시하고 계획적으로 도서를 구매하고 온라인 서점을 통해 책의 정보를 검색하는 집단이다. 군집 6은 정보비교 및 서비스 추구 형으로 온라인 서점을 통해 좀 더 다양한 서비스와 정보를 요구하는 집단이다.

군집 5(가격의존 형): 군집 5에 속한 이용자들은 주로 비용을 절약하기 위해서 온라인 서점을 이용하며 가격비교 및 할인 쿠폰 등에 관련된 서비스를 요구하는 집단이다. [그림 13]을 통해 각 군집들의 분포를 확인할 수 있다.

본 연구에서는 온라인 서점 이용자들을 의미 있는 고객 집단으로 분류하기 위해서 자율 신경망 기법 중에 하나인 자기 조직화 지도 SOM을 이용하여 고객들을 6개의 집단으로 분류하였고 기존 K-means 클러스터링 방법보다 향상된 클러스터링 결과를 가지는 것을 확인하였다. 그러나 설문지에 사용된 기본적인 인구통계학적 변수들은 각 군집에 속한 고객들의 특성을 나타내는

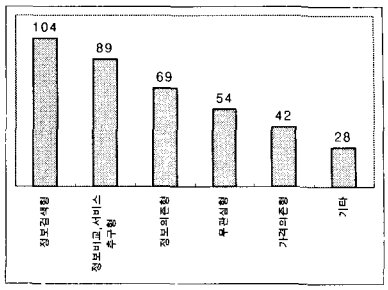



그림 13. 군집 분포

데에 한계가 있는 것으로 확인되었다. 따라서 향후 연구에서는 제안된 클러스터링 방법과 고객들의 거래 정보를 포함하고 있는 DB자료를 이용하여 각 군집에 속한 고객들의 특성 및 판매 도서들과의 관련성을 비교할 수 있는 추가 연구가 필요하다[17].

## 참 고 문 헌

[1] Pollack and Barry, "The State of Internet Marketing," Direct Marketing, Vol.61, No.9, pp.18-21, 1999.

[2] 문형남, "온라인 쇼핑몰의 고객관계관리(CRM)를 위한 소비자 구매행태 분석에 관한 연구", 한국전자상거래학회, 전자상거래학회지, 제2권, 제1호, pp.59-81, 2001.

[3] http://www.kosis.kr/

[4] 정현욱, 강혜영, 김선남, "인터넷 서점 이용자의 주관성에 관한 Q분석", 한국도서관, 정보학회지, 제36권, 제2호, 2005.

[5] J. T. Plummer, "The Concept and Application of Life Style Segmentation," Journal of Marketing, Vol.38, No.1, pp.33-37, 1974.

[6] 강병서, 김계수, *사회과학 통계분석*, SPSS 아카데미, 2001.

[7] J. Nunnally, *Psychometric Theory, second ed.*, McGraw-Hill, New York, 1978.

[8] 안광식, *교육통계방법*, 인터비젼, 2006.

[9] 김연형, 김재훈, 이석원, 이강태, *고객관계관리와*

*데이터마이닝*, 교우사, 2006.

[10] Sandhya Samarasinghe, *Neural Networks for Applied Sciences and Engineering*, Auerbach Publications, 2006.

[11] D. L. Davies and D. W. Bouldin, "Cluster Separation Measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.1, No.2, pp.95-104, 1979.

[12] T. Kohonen, "Self-Organizing Formation of Topologically Correct Feature Maps," Biological Cybernetics, Vol.43, No.1, pp.59-69, 1982.

[13] 한학용, *패턴인식 개론*, 한빛미디어, 2005.

[14] S. V. Verdu, M. O. Garcia, and C. Senabre, "Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps," IEEE Transactions on Power Systems, Vol.21, No.4, 2006.

[15] H. L. Garcia and I. M. Gonzalez, "Self-Organizing Map and Clustering for Wastewater Treatment Monitoring," Engineering Applications of Artificial Intelligence, Vol.17, No.3, pp.215-225, 2004.

[16] R. Sastry and H. Schwender, "Statistical Analysis of Occupational Safety Data of Voluntary Protection Program (VPP) and Non-VPP Sites," U.S Department of Energy, 2005.

[17] C. Perlich and Z. Huang, "Relational Learning for Customer Relationship Management," International Workshop on Customer Relationship Management. Data Mining Meets Marketing, 2005.

## 저 자 소 개

**전 현 치(Hyun-Chi Jeon)** 준회원

- 2006년 2월 : 충주대학교 산업경영공학과 (공학사)
- 2006년 2월 ~ 현재 : 고려대학교 정보경영공학부 석사과정

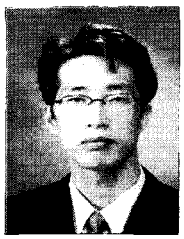
<관심분야> : CRM, 데이터 마이닝, 패턴인식

**신 영 근(Young-Geun Shin)** 준회원

- 2005년 2월 : 고려대학교 산업시스템정보공학과 (공학사)
- 2005년 9월 ~ 현재 : 고려대학교 정보경영공학부 석 박사 통합과정

<관심분야> : 패턴인식, 스케줄링, 인공지능

**박 상 성(Sang-Sung Park)** 정회원



- 2006년 2월 : 고려대학교 산업시스템정보공학과 (공학박사)
- 2006년 5월 ~ 현재 : 고려대학교 BK21 사업단 연구교수

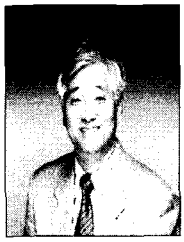
<관심분야> : 컴퓨터 비전, 패턴인식, 전문가시스템 응용, 지식관리

**김 명 훈(Myoung-Hoon Kim)** 정회원

- 2007년 2월 : 건국대학교 산업공학과 (공학박사)
- 2007년 3월 ~ 현재 : 건국대학교 강사

<관심분야> : e-비즈니스, SCM, 전문가시스템 응용, 지식관리

**장 동 식(Dong-Sik Jang)** 정회원



- 1979년 : 고려대학교 산업공학과 (공학사)
- 1985년 : 텍사스 주립대학 산업공학과 (공학석사)
- 1988년 : 텍사스 A&M 산업공학과 (공학박사)
- 1989년 ~ 현재 : 고려대학교 정보경영공학부 교수

<관심분야> : Computer Vision, 최적화이론, 컴퓨터 알고리즘