# 머신러닝(Machine Learning)을 통한 인터넷사용자 구매예측

– Entropy 의 변화량 활용 –

연세대학교 대학원

경영학과

김 민 성

# 머신러닝(Machine Learning)을 통한 인터넷사용자 구매예측

- Entropy 의 변화량 활용 -

지도교수 임 일

이 논문을 석사 학위논문으로 제출함

2015 년 1 월

연세대학교 대학원

경영학과

김 민 성

# 김민성의 석사 학위논문을 인준함

심사위원 임 일 인

심사위원 서 길 수 인

심사위원 송 민 인

연세대학교 대학원

2015 년 1 월

# 감사의 글

이제 돌아보니 대학원에서 보낸 2년이라는 시간이 정말 빨리 지나간 것 같습니다. 생각만으로도 가슴 벅찬 연구주제들을 마음껏 토론하며, 공부에만 몰입할 수 있었던 행복한 시간이었습니다.

그 동안 학문에 열정을 불태울 수 있도록 아낌없는 격려와 지도를 해주셨던 지도 교수님이신 임일 교수님께 깊은 감사의 말씀을 드립니다. 그리고 논문의 부심을 맡아 크나큰 도움을 주신 서길수 교수님과 송민 교수님께도 감사 드립니다. 또한 훌륭한 강의와 가르침을 주신 임건신 교수님, 이호근교수님, 손재열 교수님, 서응교 교수님께도 감사 드립니다. 그리고 함께 연구주제를 고민해주시고, 많은 격려와 용기를 북돋아 주신 김용학 교수님, 민순홍 교수님께 감사 드립니다.

쉽지만은 않았던 대학원 생활 동안 많은 조언을 해주신 성원언니, 곽현언니, 영은언니, 함께 공부하면서 울고 웃었던 동기 수언이, 진선씨, 혜빈, 상욱, 의곤, 든든한 후배 정원이가 있어 행복했습니다. 고맙습니다.

마지막으로 많은 지지와 격려를 해주신 부모님과 학문에 매진할 수 있도록 채찍질해준 언니, 동생 민수에게 감사의 마음을 전하고 싶습니다.

미약하지만 열심히 땀 흘려 만든 저의 이 작은 결실이 우리 학문의 발전과 전자상거래 산업의 발전에 일조할 수 있기를 간절히 바랍니다.

2015년 1월

김민성 올림

# 목 차

# 표 목차

# 그림 목차

# 국 문 요 약

## 머신러닝(Machine Learning)을 통한
## 인터넷사용자 구매예측: Entropy 변화량 활용

최근 5년동안 폭발적인 데이터양의 증가와 이를 실시간으로 처리할 수 있는 솔루션의 발전으로 빅데이터 분석은 모든 산업분야의 핵심역량이 되고 있다. 이러한 빅데이터 분석은 특히 인터넷 고객관계 관리(eCRM)에서 많이 활용되고 있는데, 전자상거래 기업들은 소비자들의 구매와 검색 기록 외에 각종 로그데이터까지 분석하여 개인화(personalization)에 활용하고자 한다. 개인화를 통한 정확한 추천시스템의 구축은 전자상거래(e-commerce) 기업의 성과를 좌우하는 중요한 요소이기 때문이다.

본 연구는 이러한 로그데이터 중에서도 클릭스트림 데이터(Clickstream Data) 분석을 통해 실시간 인터넷사용자 구매예측 모델링을 제시한다. 예측모델링에는 랜덤 포레스트(Random Forest)기법을 활용하였고, 정보 엔트로피를 지수로 활용하였다. 엔트로피 추출과 모델링은 빅데이터 분석도구로 잘 알려져 있는 오픈소스, R에서 진행하였다. 생성한 예측모델의 성과를 검증하기 위하여 국내 인터넷 마케팅 솔루션 기업이 사용자 동의아래 수집한 클릭스트림 데이터를 사용하였다. 본 데이터는 2014년 4월부터 3개월간 동일한 사용자의 PC와 모바일 사용

로그데이터와 설문조사 결과를 포함하고 있다. 검증 결과, 사용자들이 구매한 세 가지 제품군에서 각기 다른 구매패턴과 예측력이 나왔으며 최고 66%까지 예측력이 증가됨을 확인하였다.

본 연구는 빅데이터 분석을 통한 실시간 구매예측 알고리즘을 제안하고, 성과를 실증적으로 분석함으로써 빅데이터 분석의 유용성을 검증하였다는 점에서 학술적 의의를 가진다. 또한, 전자상거래 사업자들이 개인화된 배너광고 등을 통한 추천서비스 등 인터넷 고객관리 서비스를 제공함에 있어 전략적 시사점을 제공하였다는 점에서 실무적 의의를 가진다.

---

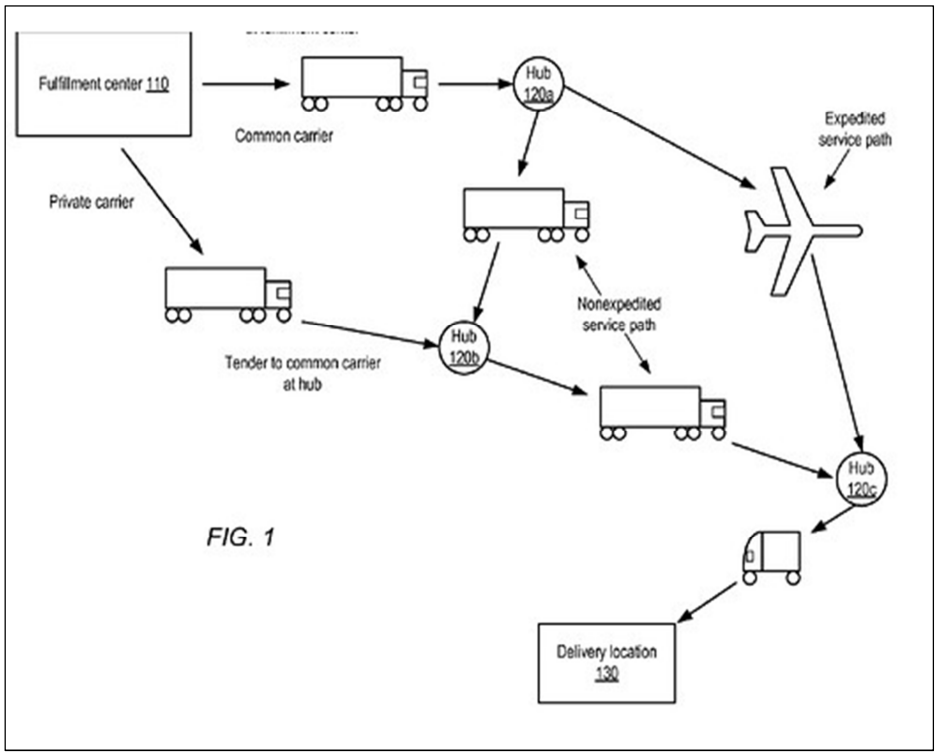
주제어: 빅데이터, 구매예측, 머신러닝, 엔트로피

# 제1장 서 론

## 1.1 연구의 배경 및 목적

최근 각종 모바일 디바이스와 소셜 미디어의 확산에 따른 실시간, 비정형 데이터가 폭증하게 되었고, 정보기술의 발달로 이를 저장하고 분석할 수 있게 되었다. 이러한 빅데이터의 분석은 다양한 형태의 데이터를 소스로 사용하여 숨겨져 있던 소비자 행동패턴을 밝히고, 일대일 마케팅(One-to-one marketing) 넘어선 실시간 예측을 통한 개인화된 서비스가 가능하도록 하였다.

이에 인터넷 고객관계 관리(eCRM)에 활용하고자 하는 기업들의 움직임도 점점 활발해지고 있다. 국내외 전자상거래 기업들은 온라인상에서 사용자들의 제품검색, 구매이력, 장바구니에 담긴 제품 정보뿐 아니라 로그데이터까지 수집하여 구매를 예측하고 이를 통해 화면에 광고를 개인화된 광고를 푸쉬(push)하여 주거나 쿠폰을 보내는 등의 고객관리에 활용하고자 한다. 가장 대표적인 전자상거래(e-commerce)기업인 아마존(Amazon)에서는 마우스의 움직임 정보 등과 같은 로그데이터까지 활용한 빅데이터 분석으로 고객이 구매결정버튼을 누르기 전에 배송을 개시하는 내용에 관련된 특허를 출원하였다. <그림1>은 이러한 아마존의 예측배송 특허문서의 일부이다.




<그림 1> 아마존 예측배송 특허 문서

빅데이터 분석을 통한 데이터 마이닝(Data Mining) 기법뿐 아니라 소비자 행동 분석에 관한 연구도 함께 활발히 이루어지고 있다.

이처럼 빅데이터 분석을 통한 데이터의 숨겨진 의미파악과 소비자 구매예측은 학문적으로나 실무적으로 화두로 떠올랐지만, 실제 기업데이터를 사용하여 데이터 특성을 고려한 측정지표를 제시하고 이를 통한 실시간 예측모델을 생성하여 실증적으로 분석한 연구는 부족한 실정이다.

이에 본 연구는 인터넷 사용자들의 구매예측을 위하여 클릭스트림 데이터를 활용하여 엔트로피 지수를 추출하였고, 이를 통해 실시간 예측모델을 생성하고 이의 성과를 검증하고자 하였다.

## 1.2 연구의 범위 및 구성

본 연구는 빅데이터를 활용한 머신러닝(Machine Learning) 기법을 사용하여, 온라인 전자상거래에서 제품 군에 따른 소비자 구매패턴을 파악하고 그에 따른 예측모델링과 성과에 대해 분석한다.

본 연구에서는 사용자 식별 아이디인 UID에 근거하여 개별 소비자의 인터넷 검색 주소인 URL과 그 검색 시간인 타임스탬프 정보가 포함된 클릭스트림 데이터와 UID에 따른 구매여부와 구매제품에 대한 정보가 포함된 설문데이터를 활용하였다. 설문데이터를 바탕으로 구매시기와 구매제품 군에 따른 소비자 UID를 추출하였고, 이렇게 추출된 UID에 해당하는 URL을 다시 날짜에 각 개별 사용자에 따른 URL 방문확률로 산출하였다. 이에 따라 소비자의 정보탐색 과정 단위로, 날짜에 따라 바뀌는 개인의 각 URL방문확률을 Pi로 활용하는 엔트로피가 결정되었다.

본 논문의 구성은 다음과 같다. 2장에서는 본 논문의 이론적 배경에 대해 기술한다. 2.1 절에서는 빅데이터의 정의 및 특성과 분석기법에 대해 기술하고, 2.2절에서는 새논(Shannon)의 정보이론(Information Theory)을

살펴본다. 2.3절과 2.4절에서는 에서는 소비자 구매의사결정 과정과 인터넷 사용자 구매패턴(구매 전 결정유보)에 관하여 기술한다. 3장에서는 섀논의 정보이론을 바탕으로 이 이론에서 나온 정보엔트로피가 본 연구에서 측정지표로 활용된 바를 기술하고, 랜덤 포레스트 기법을 통해 도출된 가설을 기술한다. 4장에서는 데이터 수집과 분석방법, 알고리즘에 관해 기술하고, 5장에서는 예측모델의 정분류율과 랜덤정분류율을 비교하여 본 연구모델의 성과를 분석한 결과를 기술한다. 마지막으로 6장에서는 결론과 향후 연구 방향에 대하여 기술한다.

# 제2장 연구의 이론적 배경

본 장에서는 본 연구환경의 토대가 되는 빅데이터와 인터넷 사용자 구매예측에 대한 이론에 대해 소개하고 이와 관련된 기존의 연구에 대해 요약하여 정리하고자 한다. 그리고 본 연구에서 측정지표로서 사용한 엔트로피와 이의 이론적 배경이 되는 섀논의 정보이론에 대한 개념을 정리하고자 한다.

## 2.1 빅데이터

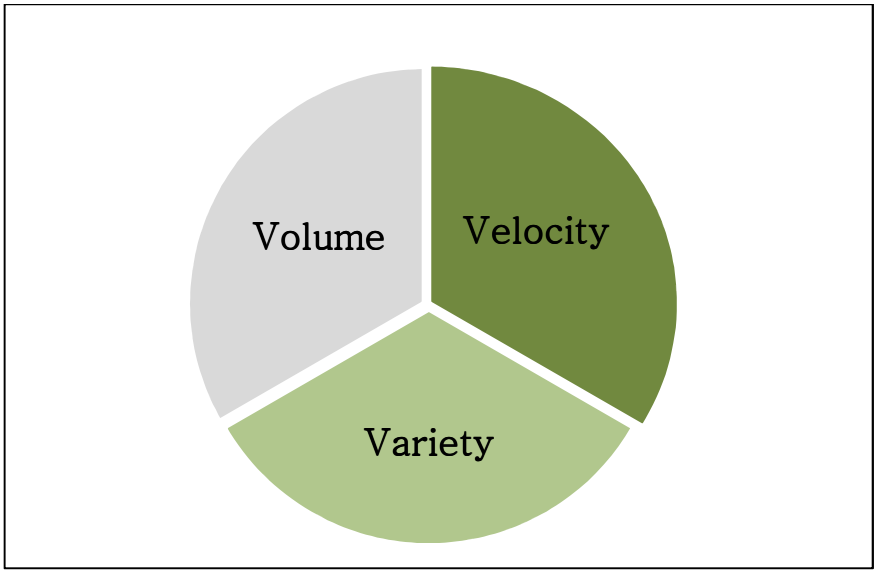
### 2.1.1 빅데이터의 정의 및 특성

2011년, 가트너는 빅데이터를 21세기의 원유에 비하며 다양한 형태의 데이터가 빠르게 생성되는 현상으로 정의하였다. 이후, 여러 기관에서 다양한 관점에서 빅데이터에 대해 정의하였고, 데이터 자체 뿐 아니라 데이터를 넘어선 데이터 분석가와 장비까지 포함시켜야 한다는 논의도 있다.

한편, Madden(2012)은 너무 크고 빠를 뿐 아니라, 고속으로 처리해야 하여 현존하는 프로세스 툴로는 처리하기 힘들 정도의 데이터라고 하였다. 데이터베이스연구자로서 그는 빅데이터 문제를 해결하기 위해서는 DB에 많은 양의 데이터를 저장하는 것만이 능사가 아니며, 통계분석과 머신러닝

알고리즘뿐 아니라 데이터 관리 에코시스템에 대한 연구와 발전이 필요하다고 주장하였다.

결론적으로 각 정의들의 공통점을 종합하면, 빅데이터는 3V(Volume, Velocity, Variety)를 갖춘, 즉 거대하고 빠르게 생성되고 처리되는 다양한 형태의 데이터로서 새로운 처리 방법이 필요하다는 것이다.




<그림 2> 빅데이터의 3가지 주요특성

데이터의 양적인 증가를 의미하는 규모(Volume)와 실시간 분석에 대한 요구의 증가를 의미하는 속도(Velocity), 텍스트나 동영상 등의 비정형 데이터까지 포함하는 다양함(Variety)외에도 4번째 주요특성으로 가치(Value)를 포함시키기도 한다(김지숙, 2012). SAS의 이러한 4V

빅데이터 모델 외에도, Gartner에서는 위 3V와 함께 전반적인 데이터 복잡도가 증가하였다고 하여 복잡성(Complexity)을 4번째 특성으로 빅데이터를 설명하였다.

### 2.1.2 빅데이터 활용 및 분석 기법

빅데이터의 등장으로 기업들은 대용량의 데이터에서 인사이트를 찾아 마케팅과 고객관리(CRM)에 활용하기 시작했다. 정치가들은 선거활동을 위한 여론분석에 이용하기 시작하였으며, 정부의 각 공공기관에서도 정책결정을 위한 자료로 빅데이터를 적극 활용하게 되었다. 그러나 거대한 데이터에서 각 분야에 필요한 인사이트를 찾는 작업은 거대한 모래사장에서 바늘을 줍는 일에 비유되기도 한다. 다양한 형태의 데이터를 전처리하는 과정 또한 녹록치 않은데 데이터 정제, 데이터 통합, 데이터 변환, 데이터 축소 등이 포함된다(윤형기, 2013).

넓게 보자면, 오늘 날 빅데이터를 처리하는 분석 기법들은 이전에도 존재하였던 것들이다. 그러나 엄청난 양으로 쏟아지는 데이터를 저장하는 스토리지(Storagae)의 문제해결과 실시간으로 처리가 가능하게 된 기술의 혁신으로 기존의 데이터 분석기법들이 새롭게 조명을 받게 되었다. 이전에는 수개월이 걸리던 분석작업이 실시간으로 처리가 가능하게 되었다는 것은 단지 처리시간의 문제가 아니라, 산업전반에 걸쳐 일어난

엄청난 혁명과도 같다. 실제로 빅데이터 분석을 활용한 실시간 개인화 추천과 부정탐지 및 불량재 예측 등은 모든 산업에 있어 기존의 패러다임을 바꾸어 놓는 결과를 낳았다.

빅데이터 분석은 분석기법 그 자체로만 본다면, 기존의 데이터 마이닝 기법의 발전과 기술축적의 산물로 볼 수 있는데, 데이터 마이닝은 거대한 양의 데이터로부터 의미 있는 정보를 추출하는 것으로 정의된다(Hand et al, 2001).

즉, 데이터 마이닝(Data Mining)은 많은 양의 데이터에서 숨겨진 패턴과 연관성 등의 유용한 정보를 발견하고 만들어내고, 이를 의사결정에 이용하는 것을 의미한다(Hand et al, 2001).

데이터 마이닝의 목적에는 예측(prediction)과 서술(description)이 있는데, 예측은 알려지지 않거나 미래가치를 데이터베이스상의 다른 속성들로 예측하는 것이고, 서술은 이해가능하고 요약된 방식의 데이터 집합으로 데이터에 대한 흥미로운 일반적 특성을 설명해주는 것을 말한다(Behrooz, 2004). 본 연구에서는 예측모델링(Predictive Modeling)을 위해 구매자와 비구매자를 분류(Classification)하는 기법을 사용한다. 분류는 데이터들을 미리 정의된 클래스들로 분류함으로써 모델을 찾기 위한 방법이다(Behrooz, 2004). 분류를 통해 현재 데이터를 설명하고, 데이터베이스의 각 클래스를 더 잘 이해하게 된다. 따라서 분류는 미래데이터 설명을 위한 모델을 제공한다(Duda et al.,2001).

분류를 통한 예측모델링 기법에는 <표1>에서 보는 바와 같이 베이지안(Baysian inference)(Duda et al.,2001), 신경망 분석(neural net approaches)(Lange, 1996), 의사결정나무(decision tree-based methods)(Quinlan, 1986), 유전자 알고리즘 분석(genetic algorithms-based approached)(Punch et al., 1993)이 있다.

<표 1> 예측모델링을 위한 분류 기법

| Ststistical method | 정의 |
|---|---|
| 베이지안 (Bayesian) | Naïve Bayesian 은 단순하면서도 매우 강력한 알고리즘으로, 실제로는 생길 수 없는 강한 록립가정을 포함하기 때문에 현실세계의 데이터셋을 침범하기도 함 |
| 의사결정나무 (Decision tree) | 한정된 수의 클래스로 예를 분류하는 것으로, 의사결정규칙을 트리구조로 도표화하여 예측을 수행하는 분석 기법 |
| 신경망분석 (Neural net) | 인간두뇌 자체를 모델화, 신경세포들의 구조에 기반하여 확률적으로 처리하는 블랙박스 형태의 기법 |
| 유전자알고리즘 (Genetic algorithms) | 진화이론에서와 같이 진화적인 방법으로 계산을 변형시키는 방법 |

앞에서 본 예측모델링에는 머신러닝 중 지도학습(supervised learning) 기법이 사용되는데, 머신러닝(Machine Learning)이란 컴퓨터가 주어진 데이터를 통해 스스로 학습하도록 하는 것이다.

데이터 마이닝 프로세스에는 유용한 패턴을 발견하기 위하여 머신

러닝과 통계(Statistics) 그리고 시각화 기법(visualization technique)이 사용된다. IDC는 2015년 IT 전망 보고서에서 머신러닝이 빅데이터 분석에 필수적이며 인공지능(Artificial Intelligence)의 새로운 이름이라고 해도 무방하다고 언급하였다. 엄밀히 말하면 머신러닝은 인공지능의 한 종류로서 컴퓨터가 주어진 데이터를 통해 스스로 학습하도록 하는 것을 의미한다(윤형기, 2013). 머신러닝의 핵심은 2가지로, 데이터의 표현과 이들에 대한 평가를 위한 함수가 첫 번째이고, 두 번째는 일반화(Generalization)이다. 현재의 모형이 새로운 데이터에도 그대로 적용될 수 있도록 하는 것이다.

머신러닝의 종류는 크게 아래 표와 같이 2가지로 나누어 볼 수 있다.

<표 2> 머신러닝의 종류

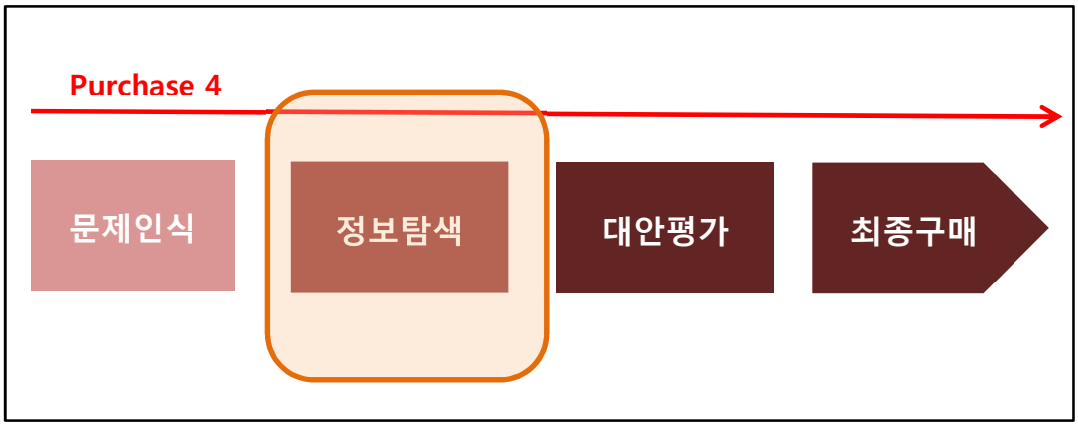
| 머신러닝(Machine Learning) | 사용되는 방법 |
|---|---|
| 지도학습<br>(Supervised learning) | 회귀(Regression)와<br>분류(Classification)의 방법이 있음 |
| 자율학습<br>(Unsupervised learning) | 군집화(Clustering)의 방법이 사용됨 |

## 2.2 섀논의 정보이론

섀논의 정보이론에 따르면 불확실성의 크기를 엔트로피로 추정할 수 있으며 이것이 사용자가 갖는 평균 정보량이 된다(SHANNON, 1948). 불확실성의 정도는 소비자의 구매 전, 정보탐색의 양을 이끌어내며 확신을 가질수록 탐색양은 줄어든다(Urbany, 1986).

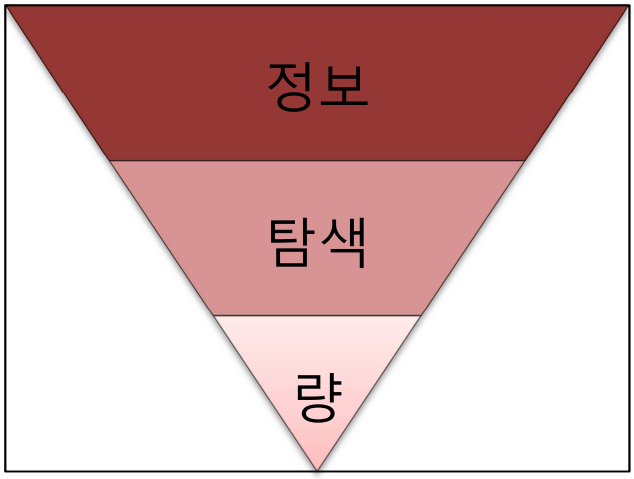
## 2.3 소비자 구매의사결정 과정

구매 전 소비자 의사결정 과정은 문제인식-> 정보탐색-> 선택 대안의 평가-> 구매 4 단계로 나누어 볼 수 있다. 이 4단계 중 정보탐색은 구매 전 발생하는 불확실성으로 위한 위험을 해소하기 위한 방법으로 널리 알려져 있다(Taloyer, 1974).




<그림 3> 소비자 의사결정 과정

이러한 불확실성의 정도는 구매 전 정보탐색의 양을 이끌어내며, 확신을 가질수록 탐색양은 줄어든다(Urbany, 1986). 구매 전, 소비자는 처음에는 모든 가능한 고려제품군(consideration set)을 스크린(screening)하다가(Fotheringham, 1988) 점차 범위를 줄여 구매고려집합(search set)에서 구매를 한다(Neveen F.Awad, 2006).




<그림 4> 고려제품군 축소에 따른 정보탐색량의 감소

이러한 행동패턴과 관련하여 Manrai and Andrews(1998)는 소비자는 제한된 정보 처리 및 획득 능력을 가지고 있어서 방문 하는 웹사이트의 수를 제한하게 된다고 하였다.

## 2.4 인터넷사용자 구매패턴(구매 전 결정유보)

2000년대에 들어서 전자상거래 (e-commerce) 의 급격한 팽창으로 시장 환경이 다양해 졌고 소비자의 선택권은 대폭 늘어났다. 그러나 점점 소비자의 선택권이 다양성을 넘어 복잡성으로 대체되었고, 소비자는 선택을 하기가 더욱 어려워진 선택의 패러독스에 빠지게 되었다. (Schwatz, 2004)

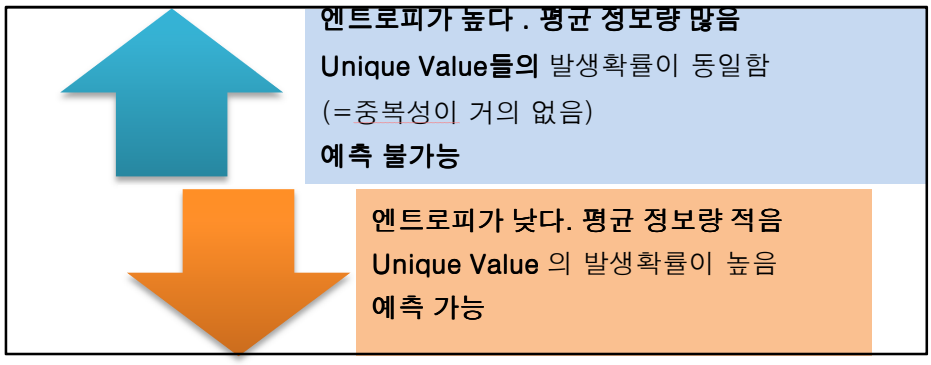
이에 따라 소비자들은 점점 결정을 유보하고 선택을 잠시 뒤로 미루는 행동을 보이기 시작하였다. Greenleaf와 Lehmann(1995)은 ‘decision delay time’을 구매욕구를 인지하고 실제로 제품을 구입하기까지의 시간이라고 정의하였다. 또한 지연의 이유에 따라 소요시간이 다르다고 분석하였다. 그들의 연구에서 분석한 지연의 이유들은 선택대안의 가능성과 추가정보획득, 또는 제품 가격이 떨어지거나 품질향상을 기다리는 것으로 보았다. Anderson(2003)의 연구에서는 ‘선택연기(choice deferral)’라는 용어를 사용하였다.

최근에 온라인 전자상거래와 관련하여 이러한 결정유보에 관한 연구들이 이루어지고 있는데 인터넷에서 제품 구매 시, 많은 사용자들이 바로 구매버튼을 클릭하지 않고 제품을 장바구니에 담아두고 더 많은 정보를 탐색한다고 밝혔다(Moore and Mathews, 2008). 또한 온라인 쇼핑에서는 제품종류에 따른 구매 행동의 차이가 존재한다는 연구들이 있다.

# 제3장 연구 가설

본 장에서는 섀논의 정보이론과 소비자의사결정이론을 기반으로, 엔트로피 지수를 활용한 예측모델과 본 기법을 적용하지 않은 분류모델의 성과를 비교하는 연구 가설을 제시한다. 또한 지수로 사용되는 엔트로피와 랜덤 포레스트 예측 모델링에 관련된 이론적 배경을 설명한다.

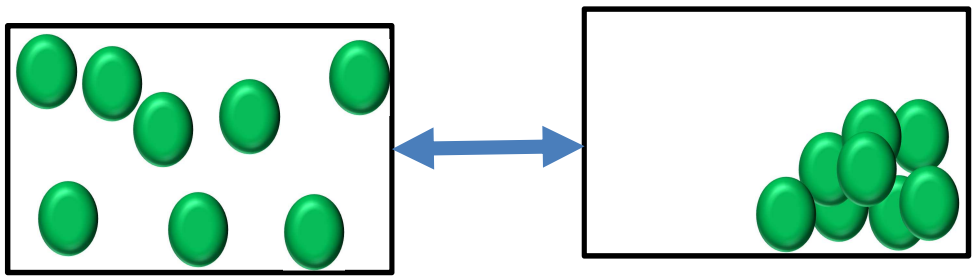
본 연구에서는 인터넷 사용자의 제품 구매 전, 정보탐색의 양이 줄어드는 것을 측정하기 위하여 엔트로피라는 개념을 활용하였다. 정보 엔트로피는 한 개의 단위정보가 감속시킬 수 있는 불확실성의 정도를 의미하는 것으로, 엔트로피가 높으면 평균정보량이 많고 unique value들의 발생확률이 동일하여 예측이 가능함을 의미한다. 반대로 엔트로피가 낮으면 평균정보량이 작고 unique value들의 발생확률이 높아 예측이 가능함을 의미한다(Chang et al., 2009).




<그림 5> 엔트로피와 평균정보량 및 예측의 관계

## 3.1 엔트로피

엔트로피(Entropy)는 측정지표로서 물리학뿐 아니라 공학, 통계학 등 다양한 분야에서 응용되고 있다.




<그림 6> 엔트로피의 고저와 무작위도의 관계

새논의 정보이론에 따르면 불확실성의 크기를 엔트로피로 측정할 수 있으므로, 소비자 구매 전 갖는 제품에 대한 불확실성의 크기는 제품탐색 과정에서 엔트로피의 감소로 나타날 것이다. 따라서 엔트로피를 지표로 이용하여 구매시점을 예측할 수 있으며, 이를 통한 예측모델링이 가능할 것으로 생각할 수 있다. 모델링 알고리즘과, 분류력을 검증하는 시뮬레이션에 관한 자세한 설명을 다음 장에 기술한다.

새논의 정보이론에 따라 엔트로피를 구하는 식은 아래와 같다.

$$E(p) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

## 3.2 랜덤 포레스트

본 연구에서는 의사결정나무의 메타학습 형태인 랜덤 포레스트 기법을 적용한 예측 모델링을 제시하였다.

랜덤 포레스트(Random Forest)는 다수의 의사결정나무를 형성하고 각각의 예측값들을 조합하여 정밀도가 높은 분류기를 얻는 기법이다(Breiman, 2001).

이때, 각각의 의사결정나무는 전체 데이터로부터 랜덤하게 선택된 일부의 변인과 표본을 이용하여 형성하므로 예측 성능이 낮을 수 있지만, 많은 수의 의사결정나무들을 조합함으로써 잡음이 많은 데이터에 대해서도 좋은 예측성능을 보이는 특징이 있다.

Random Forest방법은 Training Set에서 복원 추출 방법으로 모형을 만드는 데이터를 n개를 선택한다. 그리고 J회의 Model Set을 만들기 위하여 J번 반복을 실행하며, 각각의 실행된 횟수마다 Tree를 만든다. 그리고 예측은 각각의 Tree로부터 생성된 예측력의 평균으로 각각의 Class를 추정하는 방법이다.

따라서 이러한 논의를 바탕으로 본 연구에서는 아래의 가설을 도출하였다.

***귀무가설: 예측모델의 분류력이 랜덤예측과 차이가 없다.***

***대립가설: 예측모델의 분류력이 랜덤예측보다 정확하다.***

# 제4장 연구 방법

## 4.1 연구대상 및 데이터수집

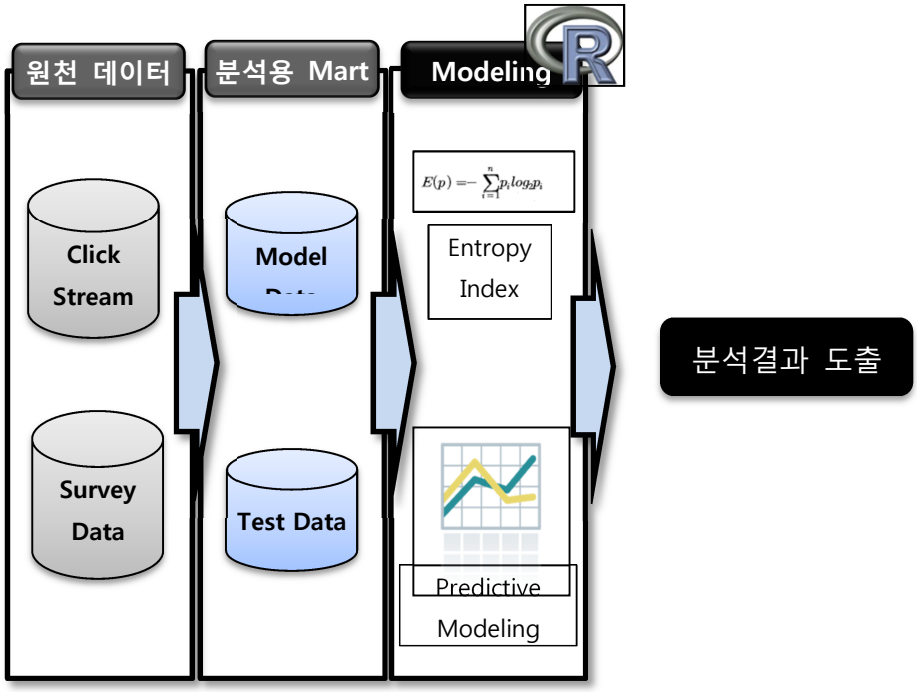
본 논문은 온라인 사용자들의 클릭스트림 데이터를 활용하여 제품군 별 구매 패턴의 차이를 살펴보고, 실시간 예측 모델 생성 및 검증을 하고자 한다. 이를 위해, 인터넷 마케팅 솔루션 기업이 사용자 동의아래 수집한 클릭스트림 데이터를 사용하였다. 데이터는 2014년 4월 1일부터 6월 30일까지 약 3개월 간의 인터넷 사용자들의 클릭스트림 데이터와 제품의 종류와 구매시기에 관해 응답한 설문조사 데이터로 구성되어 있다. 설문에 응답한 1066명에 대한 약 1억건의 로그로서 컴퓨터에서 수집된 URL은 77,868,173 건, 모바일은 7,172,025건으로 전체 약 1억건의 데이터를 분석하였다.

## 4.2 데이터의 분석도구 및 분석방법

실험을 위하여 CUBRID 와 R의 randomForest와 reshape2 패키지를 사용하여 분석 데이터셋을 생성하고, 예측모델을 만들어서 시뮬레이션을 진행하였다. 예측모델의 적용 단계에서는 일반적으로 많이 사용하는 랜덤 포레스트 기법을 사용하였다.

또한 본 연구에서는 예측모델 생성을 위해 정보 엔트로피(entropy)를

지수로 활용하였으며 이를 구하기 위하여 R에서 자체적으로 구현한 함수를 사용하였다. 자세한 순서는 아래와 같다.




<그림 7>분석 시스템 구성 및 과정

## 4.3 알고리즘

### 4.3.1 DataSet 생성

본 연구의 실험을 위하여 컴퓨터와 모바일의 클릭스트림 데이터를 사용자 식별정보(UID)로 합치고 제품 구매시기를 기준으로 엔트로피를

생성하여 통합데이터셋을 생성하였다. 데이터의 제품군은 세 가지로 나누어 살펴보았다. 구체적으로, 화장품, 패션잡화, 전자제품을 설정하였다. 각각의 제품군의 구매패턴을 분석하기 위하여 6월 중순에 구매하였고, 인터넷으로 정보를 수집한 사용자에 관한 정보를 추출하였다.

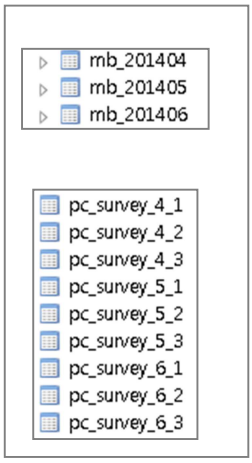
화장품 제품군을 구매한 사용자는 총 66명으로 해당 URL은 15,956,025 건으로, 약 1천6백만건이 추출되었다. 전자제품군을 구매한 사용자는 총 33명으로 해당 URL은 2,635,463 건이며 약 2백6십만건이 추출되었다. 예측모델 생성을 위하여 아무것도 구매하지 않은 사용자를 추출하였는데 이들은 총 36명으로 해당 URL리스트 2,512,862 건이 분석에 사용되었다.

Step 1: CURID DB의 설문테이블에서 타겟 사용자 UID와 제품 구매시기 및 정보수집 경로를 추출한다.

```
CREATE TABLE user_buy_6_1_2_cosmetic_search_all
AS SELECT uid, a4_2_1, a4_2_2
FROM survey
WHERE (a3_1 = 3) AND (a4_2_1 = 3 AND (a4_2_2 = 1 OR a4_2_2= 2))
AND(a4_5= 4);
```

Step 2: 위의 과정에서 추출한 타겟 사용자의 UID가 포함된 테이블과 URL리스트가 있는 클릭스트림 데이터 테이블을 조인(join)하여 타겟 사용자들에 해당하는 URL리스트 데이터를 추출한다.

| NO | uid | a4_2_1 | a4_2_2 |
|---|---|---|---|
| 1 | 20140505A942FB1A3F33 | 3 | 1 |
| 2 | 20140307E3CE872C22FF | 3 | 2 |
| 3 | 20131105C482BD163BA0 | 3 | 1 |
| 4 | 20131022B7DAB198F253 | 3 | 2 |
| 5 | 201402077D9770E57642 | 3 | 2 |




Step 3: 쿼리결과 나온 테이블을 R에서 로드한다.

Step 4: 사용자별 정보 엔트로피를 구한다. 본 연구에서는 구매 전, 정보탐색의 양이 줄어드는 것을 측정하기 위하여 정보 엔트로피를 측정지표로 활용하였다.

$$E(p) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

i: 각 URL

Pi: 각 URL의 방문확률

R에서 아래의 함수를 사용하여 사람과 날짜 별로 엔트로피를 모두 구하여 더해준다. 날짜에 따라 바뀌는 사용자들의 각 URL 방문확률을 구할

수 있다.

```
getEntropy2<-function(data){
  return(aggregate(URL~UID+date2,data=data,
                   function(x){
                     ta<-table(x)
                     p<-ta/sum(ta)
                     p2<-p[p!=0]
                     return(sum(-p2*log(p2)))
                   }))
}
ent_t<-getEntropy2(data)
```

제품 구매자와 제품 비 구매자의 엔트로피는 아래와 같이 구해지며, 예측모델에서 사용하기 위하여 Target 1을 구매자로 하고 Target0을 비 구매자로 하여 칼럼(column)을 추가하였다.

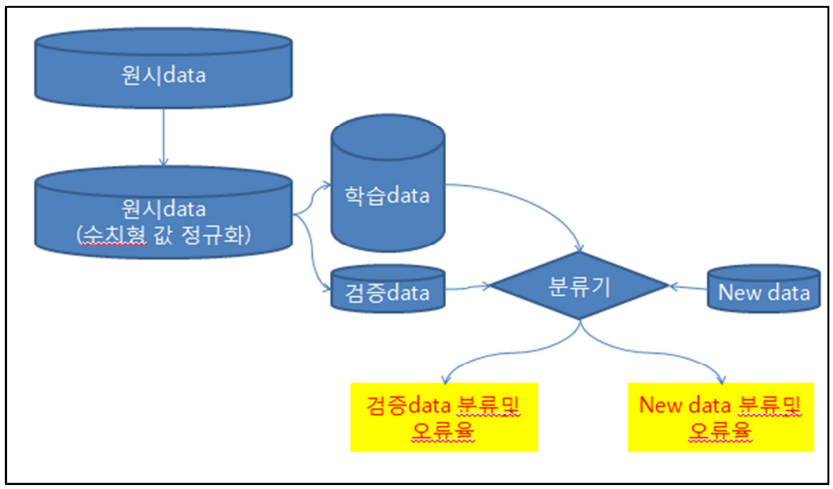
| > | model_data | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | X0 (T) | X1 | X2 | X3 | X4 | X5 | X6 | X7(T-7) | **target** |
| 1 | 7.377533 | 7.192343 | 6.886768 | 7.010467 | 6.497301 | 7.157231 | 7.29089 | 6.676898 | **1** |
| 2 | 3.209045 | 4.106475 | 3.234869 | 6.32361 | 4.368932 | 6.896528 | 6.285387 | 7.362126 | **1** |
| 3 | 5.401263 | 6.192441 | 6.587059 | 6.565418 | 6.464482 | 5.732012 | 6.129243 | 6.082805 | **1** |
| 4 | 6.128106 | 4.840796 | 6.167554 | 6.215174 | 6.216096 | 6.000465 | 5.810154 | 6.991738 | **1** |
| 5 | 0 | 0 | 0 | 6.213408 | 5.948327 | 5.91606 | 4.637926 | 5.761245 | **1** |
| 6 | 5.083311 | 5.77281 | 5.009669 | 5.496713 | 5.556627 | 5.43734 | 5.659854 | 6.360638 | **1** |
| 7 | 0 | 0 | 0 | 6.314161 | 6.839494 | 7.247681 | 6.955742 | 6.864425 | **1** |
| . | … | … | … | … | … | … | … | … | . |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 77 | 5.968431 | 5.960998 | 5.959724 | 5.849073 | 6.007657 | 5.992584 | 6.443309 | 7.747179 | **0** |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 78 | 5.948768 | 5.496158 | 6.04239 | 5.872468 | 6.019195 | 6.997382 | 6.940091 | 6.201255 | 0 |
| 79 | 5.753703 | 6.130514 | 6.449938 | 6.022351 | 5.854892 | 5.498293 | 6.648197 | 6.910751 | 0 |
| 80 | 6.296814 | 6.58362 | 5.811314 | 5.645393 | 5.884545 | 6.690828 | 6.970623 | 6.590848 | 0 |
| 81 | 7.321189 | 6.161207 | 6.960348 | 0.693147 | 2.833213 | 1.609438 | 2.302585 | 0.693147 | 0 |
| 82 | 6.668228 | 6.608001 | 1.098612 | 2.772589 | 2.197225 | 2.079442 | 0 | 2.639057 | 0 |
| 83 | 5.650118 | 5.653482 | 5.354402 | 6.349619 | 5.624844 | 6.176687 | 6.121903 | 5.855793 | 0 |
| 84 | 5.663747 | 5.732393 | 6.045456 | 5.868175 | 6.204279 | 6.119505 | 5.91118 | 6.258638 | 0 |
| . | … | … | … | … | … | … | … | … | . |

### 4.3.2 예측모델 생성

예측모델 생성을 위하여 랜덤 포레스트 기법을 적용하여 500개의 Tree를 사용하였다.




<그림 8>예측모델 시스템

### 4.3.3 예측모델 성능 평가방법

예측모델의 정확도를 측정하기 위하여 정분류율을 사용하여 예측력을 판단하였다.

<표3> 검증을 위한 정분류율표

| | | 예측값 | |
|---|---|---|---|
| | | 0 | 1 |
| 실제값 | 0 | a | b |
| | 1 | c | d |

$$\text{정분류율} = \frac{a + d}{a + b + c + d}$$

시뮬레이션은 10 Fold Cross Validation으로 시행횟수, 30회를 실시한다. 1회 평가 마다 각 Class는 Monte Carlo 방법으로 Random Generation하게 된다. 그러므로 여러 번 시행 후 그에 대한 평균 값을 사용한다.

# 제5장 데이터 분석 및 결과

본 장에서는 앞에서 생성한 사용자와 날짜별 엔트로피 지수로 본 그래프 및 예측 모델의 성과를 기술하고자 한다.

## 5.1 랜덤 정분류율

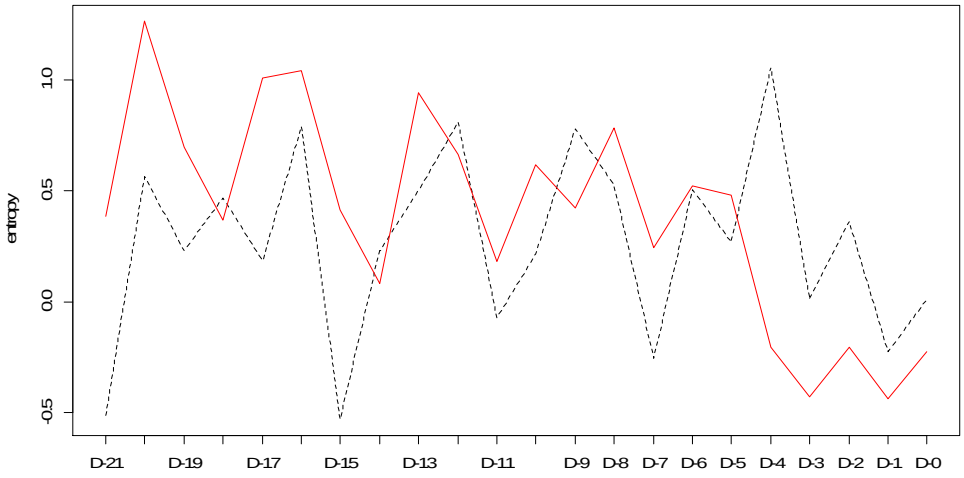
클릭스트림 데이터를 활용한 엔트로피의 정보 없이 랜덤으로 예측할 때, 이론적으로 0.5 정도의 정분류율을 얻게 된다. 이 값을 뒤의 엔트로피 정보를 활용한 예측모델의 정분류율과 비교하였다.

## 5.2 엔트로피 정보를 활용한 예측모델의 정분류율

예측모델의 시뮬레이션 결과, 제품군별로 평균값과 표준편차, 하위 95% 신뢰구간, 상위 95% 신뢰구간을 구하였다. 표본에서 구한 평균예측력은 실제 모집단을 대표하는 추정 값이나 확률적으로 그 값이 모집단 값과 정확히 일치하지는 않는다. 그러므로 우리는 확률로서 그 추정 값의 범위를 신뢰구간을 통하여 추정하여야 하며, 확률 적으로 대표 값은 신뢰구간 범위에 존재한다 라고 추정한다.

### 5.2.1 패션의류/잡화

4 장에서 생성한 데이터셋으로 제품군별로 제품구매일, D-0 인 6 월 15 일로부터 약 3 주 전인 21 일 이전까지의 엔트로피 변화를 추적하였다. 패션의류/잡화의 경우 엔트로피의 변화는 <그림 1>와 같이, 비구매집단에 비해 구매집단의 엔트로피가 구매일 1 일 전에 최저치를 기록하는 것을 볼 수 있다.




<그림 9> 패션의류/잡화군의 엔트로피 변화

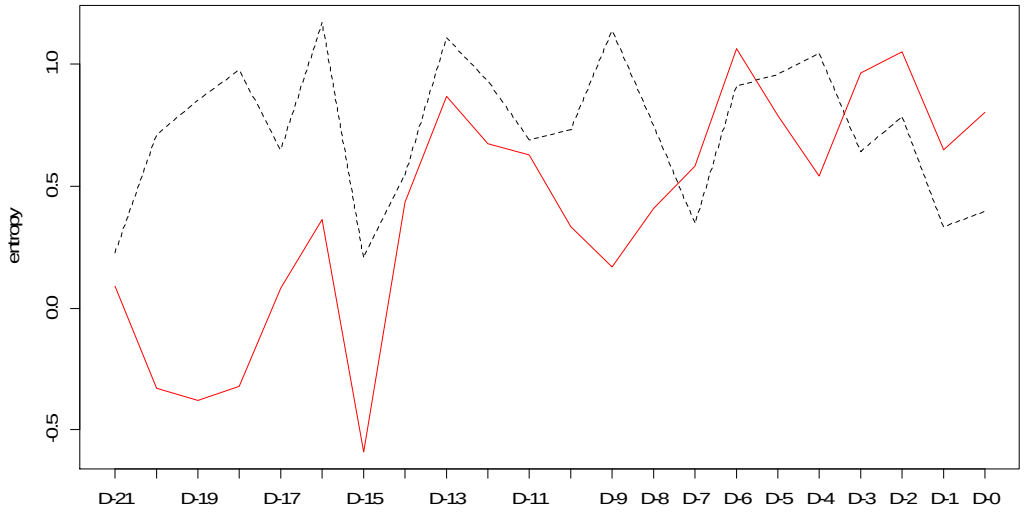
패션의류/잡화의 정분류율은 <표 1>에서 같이 95%확률로 0.82 에서 0.83 사이에 존재하고 가장 대표되는 값은 0.83 이다.

<표 4> 패션의류/잡화 제품군의 평균 정분류율, 표준편차 및 신뢰구간

| Item | Mean | SD | LCL | UCL |
|---|---|---|---|---|
| Fashion | 0.83 | 0.01 | 0.82 | 0.83 |

### 5.2.2 화장품

화장품군의 경우에는 <그림 2>와 같이 구매집단의 엔트로피가 제품구매 15 일 전에 최저치를 기록하였다.




<그림 10> 화장품군의 엔트로피 변화

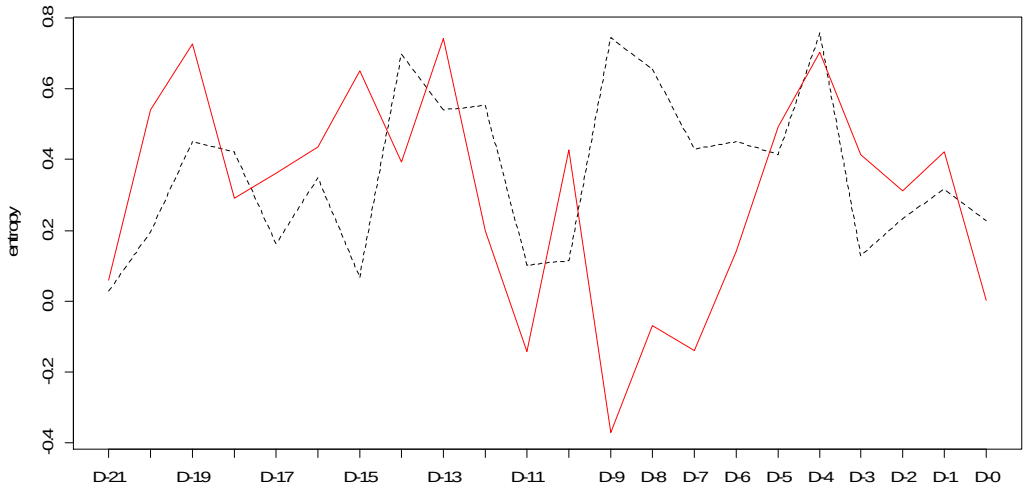
화장품군에서 예측모델의 정분류율은 패션의류/잡화의 제품군과 마찬가지로 동일한 95%확률로 0.72 에서 0.74 사이에 존재하며, 가장 대표되는 값은 0.73 이다.

<표 5> 화장품군의 평균 정분류율, 표준편차 및 신뢰구간

| Item | Mean | SD | LCL | UCL |
|---|---|---|---|---|
| 화장품군 | 0.73 | 0.02 | 0.72 | 0.74 |

### 5.2.3 가전제품

가전제품군은 구매집단의 엔트로피가 제품구매 9 일 전에 최저치를 기록하였으며, <그림 3>에서 보는 바와 같다.




<그림 11> 가전제품군의 엔트로피 변화

Electronic 의 정분류율은 95%확률로 0.54에서 0.57사이에 존재하고, 가장 대표되는 값은 0.56으로 추정된다.

<표 6> 전자제품군의 평균 정분류율, 표준편차 및 신뢰구간

| Item | Mean | SD | LCL | UCL |
|---|---|---|---|---|
| 전자제품군 | 0.56 | 0.03 | 0.54 | 0.57 |

## 5.3 랜덤과 예측모델의 정분류율 성과비교

제시된 방법의 성과를 측정하기 위하여 본 기법을 적용하지 않은 분류모델과 본 기법을 적용한 예측모델의 정분류율 10 Folder Cross Validation 방법과 Monte Carlo시뮬레이션 기법을 통하여 성능을 비교하였다.

Monte Carlo 시뮬레이션 기법은 Random Generation을 통해 난수 값을 생성하며, 그 생성된 값을 통한 추정치를 시뮬레이션 하는 기법이다. 본 연구에서는 각 10 Fold Cross Validation의 각 Class 값을 구할 때 Random Generation을 실행 하였다.

10 Fold Cross Validation기법은 머신러닝(ML)으로 모형을 생성 할 때, 생성된 모형의 예측력을 평가하는 기법이다. 먼저 전체 데이터를 10개의 그룹으로 나눈 후 첫 번째 집단을 Test Set으로 사용하고, 나머지 아홉 개의 그룹을 Training Set으로 사용하여 모델을 생성하여 Test set을 평가하게 되는데, 이러한 과정을 총 10번 시행하여 최종 모형의 성능을 평가하는 방법이다. 이 방법을 사용하면 머신러닝으로 데이터를 평가할 때 일반적으로 생기는 과적합(Overfitting)하여 예측력을 과대 추정하는 결과를 예방 할 수 있다.

Monte Carlo 시뮬레이션을 시행하여, 아래와 같은 결과를 얻을 수 있었다. 랜덤을 통한 예측력은 0.5로 볼 수 있으며, 이는 예측을 정확히 하였다와 하지

못하였다라는 사상확률이 각각 1/2이기 때문이다. 그리고 클릭스트림 데이터를 활용한 각 제품군의(패션의류/잡화군, 화장품군, 전자제품군)시간별 엔트로피 지수를 모형의 독립변수로 적용하였다. 그리고 모형은 현재 가장 머신러닝 예측 능력이 뛰어나다고 알려진 랜덤 포레스트 방법을 통하여 정분류율을 산출하였다. 그 결과는 아래와 같다.

<표 7> 제품군별 통계적 가설검정

| Item | Mean | Random | 평균 예측력 증가율 | T-값 | P-value |
|---|---|---|---|---|---|
| **패션의류/잡화** | 0.83 | 0.5 | 66% | **166.2115** | **4.89662E-46** |
| **화장품군** | 0.73 | 0.5 | 46% | **51.902** | **6.182477E-31** |
| **전자제품군** | 0.56 | 0.5 | 12% | **9.174725** | **3.26932E-10** |

<표 7>에서와 같이, 통계적 가설 검정을 통하여 T분포의 P-Value를 구하였으며, 유의수준 $\alpha = 0.05$에서 각 제품군의 P-value가 $\alpha$ 모다 매우 작다. 따라서 3장의 대립가설에서 예측하였던 대로 본 연구에서 생성한 예측모델을 사용할 때, 고객의 구매예측률을 더 높이는데 도움이 되는 것으로 나타났다.

분류력은 패션의류/잡화군이 0.83, 화장품군이 0.73, 전자제품군이 0.56으로 각각 예측률이 66%, 46%, 12% 증가되었음을 확인 하였다.

# 제6장 결론

## 6.1 연구결과의 요약 및 토의

소비자들은 제품 구매 전, 선택의 불확실성을 줄이고자 정보탐색을 실시하는데 최근 전자상거래 시장의 발달로 인한 제품의 선택폭이 넓어지고 구매 경로가 다양해지면서 소비자들의 구매패턴도 달라지게 되었다. 정보탐색 후, 결정유보의 시기가 나타나기 시작하였고 이러한 결정유보 시점에서의 정보 엔트로피의 하락은 곧 구매예측이 가능함을 시사한다. 따라서 본 연구는 클릭스트림 데이터를 활용하여 정보엔트로피를 추출하고, 이를 통한 예측모델링으로 인터넷 사용자 구매를 예측할 수 있는 방법을 개발하였다.

본 연구의 분석결과, 엔트로피 지수로 나타나는 구매패턴이 제품군마다 상이함을 확인하였다. 또한 가설 검증 결과, 엔트로피 정보를 활용하지 않은 랜덤예측보다 본 연구에서 제시하는 엔트로피 정보를 활용한 예측모델의 분류력이 더 정확한 것으로 나타났다. 패션의류/ 잡화군의 경우, 엔트로피 최저치를 기록한 1일 후에 구매가 이루어지고 랜덤 포레스트 예측모델을 적용하였을 때 예측률이 83%까지 기록하였다. 이는 엔트로피 정보 없이, 랜덤으로 예측을 하였을 경우보다 정확도가 66% 향상되었음을 보여준다.

## 6.2 연구의 시사점

### 6.2.1 연구의 이론적 시사점

본 연구는 실제 인터넷 사용자들의 로그데이터와 구매기록 데이터 분석을 통해 제품군별 구매패턴이 다르다는 것을 실증적으로 분석하였고, 정보 엔트로피 지수를 활용하여 실시간 예측 모델링을 제시한 연구이다.

빅데이터 환경에서 로그 데이터를 활용한 예측모델의 생성과 그 성과를 검증함으로써 빅데이터 분석의 유용성을 확인하였다는 점에서 학술적 의의가 있다고 할 수 있다.

또한 정보엔트로피를 활용하여 예측모델의 정확도 향상의 여지를 보여줬다는 점에서 앞으로 다른 측면에서의 정확도 향상의 밑거름이 될 수 있을 것으로 기대된다.

### 6.2.2 연구의 실무적 시사점

본 연구 결과를 토대로 전자상거래 기업 경영자의 관점에서 실무적 시사점을 살펴보면 다음과 같다.

첫째, 추후 기업들이 구매예측을 통해 개인화된 인터넷 광고를 실시함에 있어 적절한 타이밍에 Push가 가능하도록 적용할 수 있는 기법이 될

것으로 기대된다.

둘째, 서버에 쌓이는 로그데이터의 유용성을 고려하여 지금 당장은 분석에 사용하지 않더라도, 인터넷 사용자들이 남기는 각종 로그데이터를 수집하여 저장하는 시스템을 갖추어야 할 것이다. 앞으로는 데이터의 부익부 빈익빈 현상이 심화될 것이며, 되도록 많은 데이터를 수집하여 활용하는 것이 시장 우위선점의 기반이 될 것이다.

## 6.3 연구의 한계점 및 향후 연구 방향

본 연구는 구매예측 패턴을 분석하는데 있어, 클릭스트림 데이터 이외에 사용자들의 인구통계학적 정보 등 다양한 변수를 사용하지 못하였다는 한계점이 있다. 제품군 뿐 아니라, 사용자들의 나이나 성별 등도 구매패턴에 많은 영향을 미칠 것으로 예상된다.

두 번째 한계점으로 전체 클릭스트림 데이터와 예측 모델링에 활용한 사용자별 URL 로그데이터는 상당히 크지만, 제품군과 구매시기에 따라 구분한 사용자는 표본의 수가 다소 적다는 점이다.

향후 연구에서, 정보엔트로피 이외에 좀더 다양한 변수를 고려하거나 정보수집에 있어 사용자 수를 늘려서 모델링에 활용함으로써 예측모델의 성과를 좀 더 향상할 수 있을 것으로 기대된다.

# 참고 문헌


Anderson, C. J. (2003). The psychology of doing nothing: forms of decision avoidance result from reason and emotion. *Psychological bulletin, 129*(1), 139.

Awad, N. F., Jones, J. L., & Zhang, J. (2006). *Does search mater? Using online clickstream data to examine the relationship between online search and purchase behavior.* Paper presented at the Proceedings of the Twenty-Seventh International Conference on Information Systems.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Chang, C.-W., Lin, C.-T., & Wang, L.-Q. (2009). Mining the text information to optimizing the customer relationship management. *Expert Systems with Applications, 36*(2), 1433-1443.

Dhar, R., & Nowlis, S. M. (1999). The effect of time pressure on consumer choice deferral. *Journal of Consumer Research, 25*(4), 369-384.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. 2nd. *Edition. New York*.

Fotheringham, A. S. (1988). Note-Consumer Store Choice and Choice Set Definition. *Marketing Science, 7*(3), 299-310.

Gartner(2011), Big Data Analytics, *Gartner Group*

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). USING COLLABORATIVE FILTERING TO WEAVE AN INFORMATION TAPESTRY. *Communications of the ACM, 35*(12), 61-70. doi: 10.1145/138859.138867

Greenleaf, E. A., & Lehmann, D. R. (1995). Reasons for substantial delay in consumer decision making. *Journal of Consumer Research*, 186-199.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*: MIT press.

Lange, F., Hirzinger, G., Koeppe, R., Baader, A., Staudte, R., & Wei, G.-Q. (1996). Perception and manipulation in robotics: Neural network approaches *Robotics Research* (pp. 287-295): Springer.

Madden, S. (2012). From Databases to Big Data. *Ieee Internet Computing, 16*(3), 4-6.

Manrai, A. K., & Andrews, R. L. (1998). Two-stage discrete choice models for scanner panel data: An assessment of process and

assumptions. *European Journal of Operational Research, 111*(2), 193-215.

Moe, W. W. W. (2004). Dynamic Conversion Behavior at E-Commerce Sites. *Management science, 50*(3), 326-335.

Montgomery, A. L. (2004). Modeling Online Browsing and Path Analysis Using Clickstream Data. *Marketing Science, 23*(4), 579-595.

Montgomery, A. L. (2004). Predicting Online Purchase Conversion Using Web Path Analysis. *Marketing Science, 23*(4), 579.

Moore, S., & Mathews, S. (2008). An exploration of online shopping cart abandonment syndrome-a matter of risk and reputation. *Journal of Website Promotion, 2*(1-2), 71-88.

Neslin, A. (2006). Challenges and Opportunities in Multichannel Customer Management. *Journal of service research, 9*(2), 95-112.

Punch III, W. F., Goodman, E. D., Pei, M., Chia-Shun, L., Hovland, P. D., & Enbody, R. J. (1993). *Further Research on Feature Selection and Classification Using Genetic Algorithms.* Paper presented at the ICGA.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning,*

*1*(1), 81-106.

Reinartz, W. (2004). The Customer Relationship Management Process: Its Measurement and Impact on Performance. *Journal of Marketing Research, 41*(3), 293-305.

Schlosser, A. (2006). Converting Web Site Visitors into Buyers: How Web Site Investment Increases Consumer Trusting Beliefs and Online Purchase Intentions. *The Journal of marketing, 70*(2), 1-148.

Schwartz, B. (2004). The paradox ofchoice. *Why More is Less, Ecco*.

Senecal, S. (2005). Consumers'decision-making process and their online shopping behavior: a clickstream analysis. *Journal of business research, 58*(11), 1599-1608.

Shennon, K., & Weaver, W. (1948). A mathematical theory of communication. *Bell System Tehn. J, 3*, 623-637.

Sismeiro, C. (2004). Modeling Purchase Behavior at an E-Commerce Web Site: A Task Completion Approach. *Journal of Marketing Research, 41*(3), 306-323.

Stumpf, S., Bao, X., Dragunov, A., Dietterich, T. G., Herlocker, J., Johnsrude, K., . . . Shen, J. (2005). *Predicting user tasks: I know what you' re doing.* Paper presented at the 20th National

Conference on Artificial Intelligence (AAAI-05), Workshop on Human Comprehensible Machine Learning.

Urbany, J. E. (1986). An experimental examination of the economics of information. *Journal of Consumer Research*, 257-271.

van den Heuvel, C., Alison, L., & Crego, J. (2012). How Uncertainty and Accountability can Derail Strategic Save Life' Decisions in Counter-Terrorism Simulations: A Descriptive Model of Choice Deferral and Omission Bias. *Journal of Behavioral Decision Making, 25*(2), 165-187. doi: 10.1002/bdm.723

Wheeler, S. C. (2007). When the Same Prime Leads to Different Effects. *The Journal of consumer research, 34*(3), 357-368.

Wilson, I. G. (2002). So you want to get involved in e-commerce. *Industrial Marketing Management, 31*(2), 85-94.

Zhou, T., Su, R. Q., Liu, R. R., Jiang, L. L., Wang, B. H., & Zhang, Y. C. (2009). Accurate and diverse recommendations via eliminating redundant correlations. *New Journal of Physics, 11*, 19. doi: 10.1088/1367-2630/11/12/123008

고대균. (2014). 소비자의사결정과정에서의 구매고민. 서울대학교 대학원 석사학위논문

김지숙. (2012). 빅데이터 활용과 분석기술 고찰. 고려대학교 대학원 석사 학위논문

윤형기. (2013). 빅데이터: Hadoop 과 데이터 분석. 퍼플

이민우. (2009). 온라인 쇼핑에서 제품 유형과 구매 경험 및 지각된 위험이 구매 지연에 미치는 영향. 계명대학교 대학원 석사학위논문

# Abstract

## Predicting Online Purchase Using Machine Learning

Kim, Minsung

The School of Business

The Graduate School

Yonsei University

Big Data Analytics has become one of the most important technologies in e-commerce in these days. E-commerce vendors trying to apply Big Data to improve their competitive advantage by predicting purchase in internet user's behavior and hyper-personalization. It has led to a paradigm shift in the industry.

This study proposes a predictive modeling that utilizes Big Data Analytics. The online purchase was predicted by a random forest method with information entropy.

In order to test performance improvement by this predictive model,

an empirical study was conducted using clickstream data. The proposed model was compared with random predictive model. The empirical results show that performance improved about 66% when the proposed algorithm was used with particular product category (fashion & accessories).

This study has several theoretical and practical implications. This study empirically shows that Big Data Analytics can affect the performance of predictive systems. This helps researchers understand factors affecting performance of predictive modeling with Big Data. This study also opens a door for future studies in the area of applying information entropy to predictive model to analyze characteristics of dataset. In practice, this study provides guidelines to improve performance of predictive modeling systems with a simple modification.

Key Words: bigdata, predict purchase, machine learning, entropy