토픽 분석을 활용한 관심 기반 고객 세분화 방법론

현윤진* · 김남규** · 조윤호***

Interest-based Customer Segmentation Methodology Using Topic Modeling

Yoonjin Hyun* · Namgyu Kim** · Yoonho Cho***

Abstract

As the range of the customer choice becomes more diverse, the average life span of companies' products and services is becoming shorter. Most companies are striving to maximize the revenue by understanding the customer's needs and providing customized products and services. However, companies had to bear a significant burden, in terms of the time and cost involved in the process of determining each individual customer's needs. Therefore, an alternative method is employed that involves grouping the customers into different categories based on certain criteria and establishing a marketing strategy tailored for each group. In this way, customer segmentation and customer clustering are performed using demographic information and behavioral information. Demographic information included sex, age, income level, and etc., while behavioral information was usually identified indirectly through customers' purchase history and search history. However, there is a limitation regarding companies' customer behavioral information, because the information is usually obtained through the limited data provided by a customer on a company's website. This is because the pattern indicated when a customer accesses a particular site might not be representative of the general tendency of that customer. Therefore, in this study, rather than the pattern indicated through a particular site, a customer's interest is identified using that customer's access record pertaining to external news. Hence, by utilizing this method, we proposed a methodology to perform customer segmentation. In addition, by extracting the main issues through a topic analysis covering approximately 3,000 Internet news articles, the actual experiment applying customer segmentation is performed and the applicability of the proposed methodology is analyzed.

Keywords : Customer Segmentation, Data Mining, Recommendation Systems, Text Mining, Topic Modeling

논문접수일 : 2014년 12월 26일 1차 논문수정일 : 2015년 02월 10일 2차 논문수정일 : 2015년 02월 21일 논문게재확정일 : 2015년 02월 21일

※ 이 논문은 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2014S1A5A2A01010677).

* 국민대학교 비즈니스IT전문대학원 박사과정, e-mail : yoonjin0630@kookmin.ac.kr

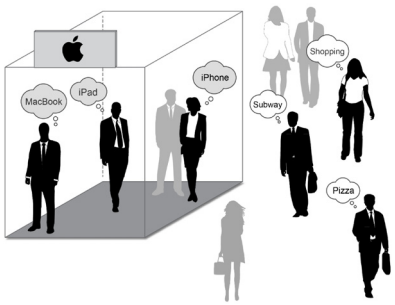
** 교신저자, 국민대학교 경영정보학부 부교수, e-mail : ngkim@kookmin.ac.kr

*** 국민대학교 경영학부 교수, e-mail : www4u@kookmin.ac.kr

1. 서 론

대부분의 기업들은 고객이 원하는 것을 파악하고 그에 맞는 맞춤형 제품과 서비스를 제공함으로써 수익을 극대화하기 위해 많은 노력을 기울이고 있다. 하지만 고객 개개인의 니즈(Needs)를 파악하는 것은 많은 어려움이 따를 뿐 아니라 시간과 비용이 지나치게 많이 소비되기 때문에, 이에 대한 대안으로 고객 세분화 즉 고객 클러스터링을 수행하게 된다. 고객 클러스터링의 본질적인 목적은 고객을 동질적인 특성을 가진 집단으로 나누어 보다 손쉽게 고객의 니즈를 파악해 차별화된 마케팅 전략을 수립하는 데에 있다. 기존의 고객 클러스터링은 일반적으로 성별, 연령, 소득 등의 단순하면서도 범용적인 정보, 즉 원시적 형태의 인구통계학 정보를 활용하여 수행되어 왔으며, 최근에는 각 기업들이 보유하고 있는 고객의 구매 내역, 검색 내역 등의 고객 행태 정보를 고객 클러스터링에 활용하고 있다[고은주 외, 2005; 이건창, 정남호, 2003; 진서훈, 2005].

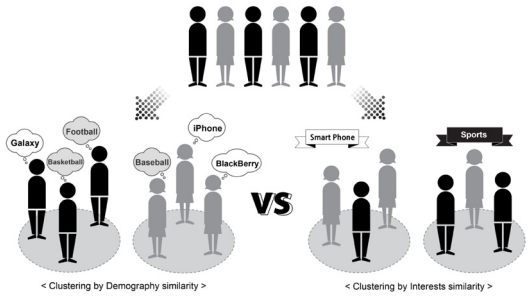
하지만 기업이 보유하고 있는 고객의 행태 정보는 해당 고객이 자사의 사이트 내에 진입한 이후에 보이는 한정된 정보만을 활용한다는 측면에서 한계를 갖는다. 예를 들어 <그림 1>에서 실제 고객들은 “Shopping”, “Subway”, “Pizza” 등 다양한 주제에 대해 관심을 갖고 있지만, 특정 기업의 사이트에 진입한 이후 고객들이 나타내는 관심



〈그림 1〉 실제 관심과 특정 사이트 내에서 표출되는 관심과의 차이

주제는 “MacBook”, “iPad”, “iPhone”으로 해당 기업과 관련된 내용으로 국한되어 나타나는 현상을 보이고 있다.

즉, 개별 기업이 보유하고 있는 고객의 행태 정보는 자사의 사이트 내에서 보이는 구매 및 검색 내역에 국한되기 때문에, 이에 기반한 고객 클러스터링은 고객의 실제 관심의 동질성이 아닌 해당 사이트에서 표출된 관심의 동질성에 기반하여 이루어지게 된다. 한편 클러스터링은 제시된 기준 측면에서의 동질성을 기반으로 수행되므로, 기준이 상이하면 클러스터링 결과 역시 상이하게 나타난다. 예를 들어 <그림 2>는 성별의 동질성 관점에서 클러스터링을 수행한 경우 서로 다른 관심을 갖는 고객들이 동일 클러스터에 소속될 수 있고, 반대로 관심 측면에서 클러스터링을 수행한 경우 서로 다른 성별의 고객들이 동일 클러스터에 소속될 수 있음을 보여주고 있다. 즉 클러스터링 기준에 따라 결과 클러스터가 상이하게 나타나므로, 클러스터링의 기준을 분석의 목적에 따라 제대로 설정하는 것이 매우 중요하다고 할 수 있다.



〈그림 2〉 클러스터링 기준에 따른 결과의 차이

따라서 본 연구에서는 고객이 특정 사이트 내에 진입한 이후에 보이는 한정된 정보만을 기준으로 활용하는 기존의 고객 세분화와 달리, 고객의 실제 관심에 기반한 고객 세분화를 수행할 수 있는 방안을 제시하고자 한다. 구체적으로는 고

객이 평소 조회한 뉴스 기사에 대해 토픽 분석을 수행하고, 이를 통해 파악된 각 고객의 관심 분야를 고객 세분화에 활용하고자 한다.

본 논문의 이후 부분은 다음과 같이 구성된다. 제 2장에서는 본 연구의 수행을 위한 핵심 기법인 텍스트 마이닝, 토픽 분석 및 고객 세분화에 대한 선행 연구들을 요약하고, 제 3장에서는 본 연구에서 제안하는 관심 기반 고객 세분화 방법론을 소개한다. 제 4장에서는 실제 수집한 데이터를 대상으로 본 제안 방법론을 적용한 실험 결과를 분석하고, 마지막 제 5장에서는 본 연구의 기여 및 한계, 그리고 향후 연구방향을 제시한다.

2. 관련 연구

2.1 텍스트 마이닝 및 토픽 분석

텍스트는 현실에서 정보를 교환하거나 표현하는 방법으로 가장 널리 사용되는 수단이다[Witten, 2004]. 최근에는 새로운 기술의 발전과 더불어 인터넷 문화가 확산됨에 따라, 웹과 다양한 소셜미디어를 통해 방대한 양의 텍스트 데이터가 더욱 활발히 유통되고 있다. 이러한 다량의 텍스트 형태의 비정형 데이터를 분석하여 이전에는 찾을 수 없었던 새롭고 의미 있는 정보를 추출하는 과정을 텍스트 마이닝이라고 할 수 있다[Hearst, 1999; Sebastiani, 2002]. 기존의 데이터 마이닝에서 확장된 형태의 텍스트 마이닝은 전통적인 데이터 마이닝에서 사용되는 연관관계 분석(Association), 분류(Classification), 군집화(Clustering)뿐만 아니라, 자연어 처리, 정보 검색, 전산 언어학, 토픽 추적(Topic Tracking), 텍스트 범주화(Text Categorization) 등 분야의 기술을 종합적으로 활용한다[Mooney and Bunescu, 2006; Stanvrianou et al., 2007]. 특히 자연어 처리 기술은 텍스트 마이닝 분석 결과의 성패를 좌우하는 핵심 기술이라고 말할 수 있다.

토픽 분석(Topic Analysis)은 텍스트 마이닝의 대표적인 분석 기법으로, 각 문서에 포함된 용어의 빈도수에 근거하여 유사 문서를 그룹화한 뒤, 각 그룹을 대표하는 주요 용어들을 추출하여 해당 그룹의 토픽 키워드 집합을 제시하는 방식으로 이루어진다. 토픽 분석은 벡터공간모델(Vector Space Model)[Salton et al., 1975; Sebastiani, 2006]과 TF-IDF(Term Frequency-Inverse Document Frequency)[Han et al., 2011; Provost and Fawcett, 2013]를 주요 이론적 배경으로 갖는다. 벡터공간모델은 텍스트를 표현하는 기본적인 방법으로, 분석 목적에 따라 행렬, 계층, 벡터 등의 다양한 형태로도 표현이 가능하며[Albright, 2006], TF-IDF는 여러 문서에서 자주 출현하는 일반적인 단어는 가중치를 낮게 부여하고, 특정 문서에만 출현하는 비일반적인 단어의 가중치는 높게 부여하는 계산 방식이다. 각 문서는 용어 수만큼의 차원과 TF-IDF를 값으로 갖는 벡터로 표현되는데, 문서 내 존재하는 용어의 수가 너무 방대하기 때문에 SVD(Singular Value Decomposition) 등을 활용한 차원 축소가 이루어지게 된다[Salton et al., 1975]. 이러한 과정을 통해 비정형 텍스트 문서의 구조화가 완료되면 이후 과정에서 정형 데이터와 함께 텍스트 데이터에 대한 군집화, 예측 등의 작업이 가능해진다.

최근에는 토픽 분석을 활용하여 이슈를 추적하는 이슈 생명주기 분석[임명수, 김남규, 2014], 개인의 관심 흐름[류신, 김남규, 2014] 등의 연구도 활발히 이루어지고 있으며, 국내의 기업들도 텍스트 형태의 비정형 데이터를 분석함으로써 기존에는 파악하지 못했던 유용한 정보를 얻어내고 있다. 특히 다음소프트의 소셜미디어 분석 솔루션인 '소셜 메트릭스' 서비스는 문맥 중심의 텍스트 마이닝 작업을 통해 각 데이터의 출현 원인 및 다른 데이터들과의 관계를 도식화하여 제공하고

있으며, 와이즈넛의 경우 트위터, 페이스북, 블로그, 카페 등에 올라온 대선 후보 관련 버즈(Buzz)를 분석하기 위한 '버즈인사이트바이럴 지수(BVI)'를 개발하였다. 또한 실시간으로 사용자의 반응과 이슈 등을 분석 보고하는 다이퀘스트의 '브람스(Brams)', 다양한 이슈나 트랜드 등에 관한 정보를 제공하는 솔트룩스의 '트루 스토리(True Story)' 등이 토픽 분석을 활용한 비정형 데이터 분석 서비스의 대표적인 예라고 할 수 있다. 토픽 분석은 유사한 주제를 갖는 문서들을 묶어서 하나의 토픽으로 구성한다는 점에서는 전통적인 군집화와 유사한 특성을 갖지만, 하나의 문서가 다수의 토픽에 대응될 수 있다는 점에서 기존의 군집화와 차별성을 갖는다. 본 논문에서는 토픽 분석과 기존의 군집화 방법을 모두 활용하여, 토픽 분석의 결과를 기반으로 고객 클러스터링을 수행한다.

2.2 고객 세분화

고객 세분화는 고객을 동질적인 특징을 갖는 집단으로 나누는 것으로 정의되며, 고객관계관리(CRM)의 핵심이라고 할 수 있다. 고객 세분화의 본질적인 목적은 소비자의 다양한 욕구를 파악하여 보다 효율적으로 기업의 마케팅 전략을 세우는 데에 있다[문권모, 2004]. 특히 기업의 타겟 마케팅, 신규고객 획득, 우수고객 유지, 고객가치 증진, 잠재고객 활성화, 평생 고객화 등의 마케팅 전략의 일환으로 주로 활용된다.

고객 세분화의 기준은 매우 다양하나 일반적으로 마케팅 분야에서 사용되고 있는 포괄적 기준은 인구통계학적 변수, 지리학적 변수, 심리적 변수, 인지 및 구매행동 변수 등을 포함한다[박명호 외, 2005; 송민영, 2001; 이두희, 2006]. 인구통계학적 변수는 나이, 성별, 연령, 직업, 소득, 가족상황 등을 의미하며, 지리적 변수는 국가, 지역, 온라인 또는 오프라인 등을 의미한다. 이러한 인구통계학적 변수는 비교적 측정이 용이할 뿐만 아니라, 각종 통계자료를 이용하여 파악할 수 있기 때문에 일반적으로 가장 널리 사용되고 있는 세분화 변수이다. 심리적 변수는 라이프 스타일, 사회계층, 성격유형 등을 의미하는데, 인구통계학적 변수에 비해 추상적이므로 측정 자체가 어렵고 세분화된 고객에 대한 접근 가능성이 낮다. 그럼에도 불구하고 심리적 변수는 소비자의 행동을 보다 근원적으로 설명해줄 수 있는 변수이다. 인지 및 행동변수는 제품 및 서비스에 대한 태도, 이용률, 충성도, 제품 및 서비스에 대한 구매패턴 등을 의미한다. 인지 및 행동변수는 고객의 구매행동을 직접적으로 나타냄과 동시에 실질적인 구매 행동과 밀접한 관련이 있는 정보를 나타내므로, 고객을 세분화 하는데 있어서 가장 효과적인 변수라고 할 수 있다.

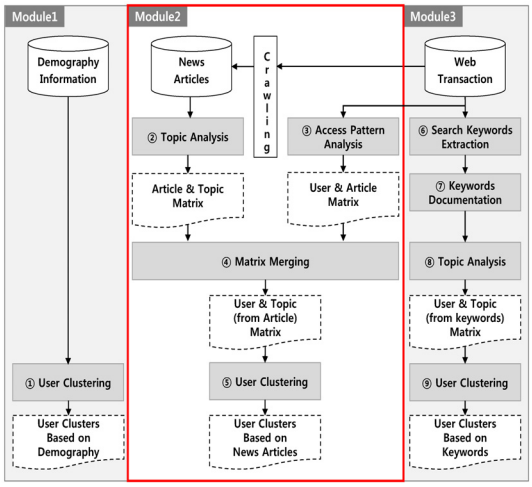
이처럼 다양한 세분화 기준을 이용하여 매우 다양한 산업에 걸쳐서 고객 세분화가 활발히 이루어지고 있다. 대표적인 예로 정책 고객의 개념에 따른 유형 분류를 통한 정책마케팅 적용 방안[김세훈, 2006], 국적 항공사의 상용고객 우대 프로그램 회원의 마일리지 자료를 통한 고객가치 산출과 고객 세분화[박광식, 2014], 대학 한방병원 통합의료정보시스템의 외래환자 인구학적 정보와 진료기록 정보 기반의 고객 세분화를 통한 외래환자 고객만족도 증진을 위한 고객관계관리 시스템 구축 모형[안요찬, 2010], 국내의 특정 철강 유통사 대상의 심층적 사례분석을 통한 고객 세분화 기준변수 및 고객 세분화 모형 개발[윤종수, 윤종욱, 2004], 이동통신회사 고객 이탈률, 고객의 현재가치, 고객의 잠재가치 측정을 통한 고객 세분화 방안[정태수 외, 2003], 풀 서비스 레스토랑을 대상으로 고객 지향적 서비스의 구성요소를 도출한 연구[한희섭 외, 2013] 등을 들 수 있다. 또한 최근에는 데이터 마이닝 기법을 활용한 고객 세분화 기법이 각광을 받고 있으며, 대표적인

기법들로는 규칙기반 시스템, 의사결정나무, 인공신경망, K-평균 군집방법 등이 주로 활용되고 있다[오은녕, 이희상, 2002; 조홍규, 2003; 정태수 외, 2003; 진서훈, 안상욱, 2004; 채경희, 김상철, 2010; Elmer, Borowski, 1988]. 이들 대부분의 연구는 특정 기관 혹은 기업 내 고객 정보 및 정형 데이터를 활용하여 고객 세분화를 수행했다는 측면에서, 외부 데이터인 고객의 뉴스 접속 기록과 포털사이트 검색 기록에 대한 비정형 토픽 분석을 통해 고객의 관심을 식별하고 이에 근거한 고객 세분화를 수행하는 본 연구와는 차이가 있다.

3. 관심 기반 고객 세분화 방법론

3.1 연구 범위

본 절에서는 토픽 분석을 활용한 관심 기반 고객 세분화 방법론을 구체적으로 제시한다. <그림 3>은 본 연구에서 제안하는 방법론의 전체 구조를 보여주고 있으며, 실험 데이터는 원통형으로, 각 단계를 구성하는 세부 프로세스는 사각형으로 표시되어 있다. 또한 중간 산출물은 점선 도형으로 나타나있다.



〈그림 3〉 전체 연구 개요

제안 방법론은 (1) 인구통계학 정보 기반의 고객 클러스터링(Module 1), (2) 뉴스 접속 기록 기반의 고객 클러스터링(Module 2), 그리고 (3) 포털사이트 검색 키워드 기반의 고객 클러스터링(Module 3)의 세 부분으로 구성되어 있다. 본 연구의 핵심은 고객이 방문한 인터넷 뉴스 기사에 대한 토픽 분석을 통해 고객의 실제 관심을 파악하고 이를 기반으로 고객 클러스터링을 수행하는 것이며, 이는 Module 2에 나타나 있다. Module 1과 Module 3은 클러스터링 기준에 따라 고객 세분화 결과가 상이하게 나타남을 보여주기 위해 부가적으로 수행된다. 주요 단계 중 토픽 분석과 클러스터링은 기존의 이론 및 상용화 도구를 통해 수행된다. 한편 매트릭스 병합은 두 가지 서로 다른 관점의 매트릭스에 대한 행렬 곱을 통해 고객의 관심 매트릭스를 도출하는 과정으로 본 연구에서 새롭게 제안되는 부분이다.

Module 1의 경우 고객의 Demography Information을 기반으로 ① User Clustering을 수행함으로써 가장 기본적인 방식의 고객 클러스터링을 수행하게 된다. 뉴스 접속 기록 기반의 고객 클러스터링인 Module 2의 경우, Web Transaction에 기록된 URL을 이용하여 News Articles를 수집하고, 수집된 News Articles에 대하여 ② Topic Analysis를 통해 다수의 토픽을 추출한 후, 뉴스 기사와 토픽간의 대응 매트릭스를 도출한다. ③ Access Pattern Analysis 단계에서는 인터넷 뉴스 사이트의 접속 기록을 분석하여 고객과 뉴스 기사 간 대응 매트릭스를 도출하고, ④ Matrix Merging 단계에서는 앞서 도출된 ②의 결과물인 뉴스 기사와 토픽 간 대응 매트릭스와 ③의 결과물인 고객과 뉴스 기사 간 대응 매트릭스를 병합하여 고객과 토픽 간의 대응 매트릭스를 도출한다. 마지막으로 ⑤ User Clustering 단계에서 ④의 결과물인 고객과 토픽 간 대응 매트릭스를 활용하여 뉴스 접속 기록 기반의 고객 클러스터링을

수행하게 된다. 포털사이트 검색 키워드 기반의 고객 클러스터링인 Module 3의 경우, ⑥ Search Keywords Extraction 단계에서 고객이 포털 사이트에서 검색한 키워드를 모두 추출한 후, ⑦ Keywords Documentation 단계에서는 검색어들을 각 고객별로 통합하여, 고객별로 하나의 키워드 문서를 생성한다. 이렇게 구성된 키워드 문서들에 대해 ⑧ Topic Analysis를 수행하여 다수의 토픽을 추출한 후, 사용자와 토픽간의 대응 매트릭스를 도출하여 ⑨ User Clustering 단계에서 검색 키워드 기반의 고객 클러스터링을 수행하게 된다. 이후 절에서는 Module 1~Module 3의 세부 과정에 대한 자세한 내용을 소개한다.

3.2 Module 1 : 인구통계학 정보 기반의 고객 클러스터링

앞에서도 언급하였듯이, Module 1의 인구통계학 정보 기반의 고객 클러스터링은 고객의 인구통계학 정보를 기반으로 고객을 그룹화 하는 가장 전통적인 고객 클러스터링 방법이다. 일반적으로 인구통계학 정보는 나이, 성별, 연령, 직업, 소득, 학력 등에 대한 정보를 의미하며, <표 1>은 특정 고객에 대한 인구통계학 정보를 기반으로 가상 클러스터링을 수행한 결과를 보여주고 있다.

〈표 1〉 인구통계학 정보 기반의 고객 클러스터링 예

Cluster No	User ID	Gender	Age	Job
1	U1	여자	22	학생
	U2	여자	25	학생
	U3	남자	27	학생
2	U4	여자	30	회사원
	U5	남자	32	회사원
	U6	남자	35	회사원

<표 1>을 보면 “U1”, “U2”, “U3” 3명의 고객이 성별은 다르지만 비슷한 연령대와 동일한 직업을 가짐으로써 하나의 고객 그룹으로 묶인 것을 알 수 있다. 이와 마찬가지로 “U4”, “U5”, “U6” 3명의 고객 역시 성별은 다르지만 비슷한 연령대와 동일한 직업을 가짐으로써 하나의 고객 그룹으로 묶인 것을 알 수 있다. 이처럼 인구통계학 정보 기반의 고객 클러스터링은 고객의 성별, 나이, 직업 등의 비슷한 정도에 따라 고객 그룹을 형성한다.

3.3 Module 2 : 뉴스 접속 기록 기반의 고객 클러스터링

Module 2의 뉴스 접속 기록 기반 고객 클러스터링은 본 연구에서 새롭게 제시하는 방법으로, 고객의 뉴스 기사 접속 기록을 분석하여 각 고객이 관심을 갖는 뉴스 토픽을 기반으로 고객을 그룹화한다. 우선, Web Transaction에서 인터넷 뉴스 사이트 접속 기록을 추출하여 크롤링을 통해 해당 News Articles를 수집한 후, 토픽 분석을 수행함으로써 다수의 토픽을 추출함과 동시에 뉴스 기사와 토픽 간 대응 매트릭스를 도출하게 된다. <표 2>는 토픽 분석으로 추출된 가상의 토픽을 보여주고 있다.

〈표 2〉 토픽 분석을 통한 뉴스 기사의 가상 토픽 추출 예

Topic ID	Topic Terms	Name	Member Articles
T1	스포츠, 선수, 야구, 야구장, 시즌	스포츠	A1, A2
T2	건강, 질병, 운동, 환자, 병원	건강	A1, A3, A5
T3	현대, 기아, 수출, 자동차, 외제차	자동차	A4, A6
T4	교육, 학원, 사교육, 선생님, 정책	교육	A4

고객의 인터넷 뉴스 기사 접속 기록을 분석하여 고객과 뉴스 기사 간 대응 매트릭스를 도출한 후, 앞서 도출된 뉴스 기사와 토픽 간 대응 매트릭스와 병합하여 고객과 토픽 간 대응 매트릭스를 도

출한다. <그림 4>는 <표 2>의 예를 활용한 매트릭스 병합 과정의 예를 보여주고 있다.

(a) Article

User	A1	A2	A3	A4	A5	A6
U1	1	1	0	0	0	0
U2	0	0	1	0	1	0
U3	0	0	1	0	1	0
U4	0	0	1	0	1	0
U5	0	0	0	1	0	1
U6	0	0	0	0	0	1

(b) Topic

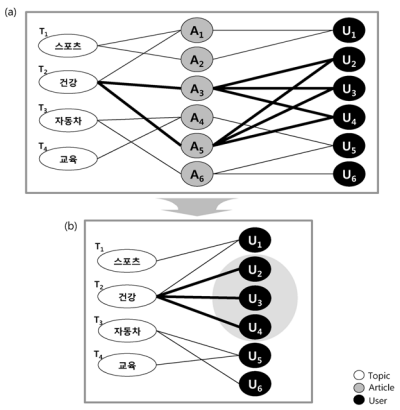
Article	T1	T2	T3	T4
A1	1	1	0	0
A2	1	0	0	0
A3	0	1	0	0
A4	0	0	1	1
A5	0	1	0	0
A6	0	0	1	0

(c) Topic

User	T1	T2	T3	T4
U1	2	1	0	0
U2	0	2	0	0
U3	0	2	0	0
U4	0	2	0	0
U5	0	0	2	1
U6	0	0	1	0

〈그림 4〉 매트릭스 병합의 예

<그림 4(a)>는 고객과 뉴스 기사 간 대응 매트릭스를 나타내고 있으며, <그림 4(b)>는 뉴스 기사와 토픽 간 대응 매트릭스를 나타내고 있다. <그림 4(c)>는 위의 두 매트릭스에 대한 행렬 곱을 통해 도출된 고객과 토픽 간 대응 매트릭스로, 각 고객이 각 토픽에 해당되는 기사 몇 건을 접속했는지의 정보를 요약해서 나타내고 있다. <그림 5>는 <그림 4>의 매트릭스를 이용하여 뉴스 접속 기록 기반의 고객 클러스터링을 수행한 결과를 네트워크 형태로 보여주고 있다.



〈그림 5〉 뉴스 접속 기록 기반의 고객 클러스터링의 예

<그림 5(a)>는 <그림 4>의 매트릭스 병합 과정을 네트워크 형태로 표현한 것으로, 이 결과를 활용하여 클러스터링을 수행한 결과는 <그림 5(b)>에 나타나있다. <그림 5(b)>를 보면 <표 1>에서 상이한 그룹에 속해있던 “U2”, “U3”, “U4” 3명의 고객이 “건강”이라는 공통된 뉴스 관심사를 가짐으로써 새로운 고객 그룹을 형성한 것을 알 수 있다. 즉, 인구통계학 정보가 아닌 관심사라는 새로운 기준으로 클러스터링을 수행함으로써 새로운 고객군을 식별할 수 있다.

3.4 Module 3 : 포털사이트 검색 키워드 기반의 고객 클러스터링

Module 3의 포털사이트 검색 키워드 기반 고객 클러스터링 역시 본 연구에서 새롭게 제시하는 방안이다. Module 2가 이미 게시된 뉴스의 조회를 통해 고객의 관심을 간접적으로 파악하는 방법이라면, Module 3은 사용자가 직접 입력한 키워드를 통해 고객의 관심을 파악한다는 점에서 더욱 직접적인 방법이라고 할 수 있다. 우선, Web Transaction에서 각 고객의 인터넷 포털사이트 검색 기록을 바탕으로 해당 고객이 입력한 검색 키워드를 모두 추출한 후, 추출된 검색 키워드들을 고객별로 나누어 하나의 문서 형태로 가공한다. 즉, 고객별로 하나의 키워드 문서를 형성하게 되며, 해당 문서는 특정 고객이 포털사이트에서 검색한 모든 키워드들의 묶음을 의미한다(<표 3> 참조).

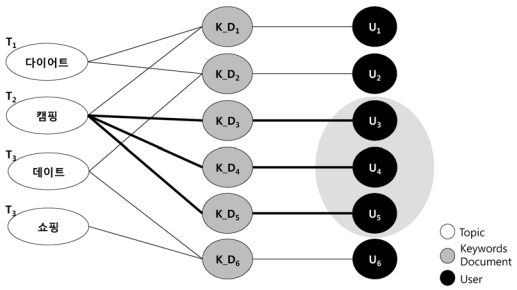
〈표 3〉 고객별 키워드 문서의 예

User ID	Keywords Document ID	Contents
U1	K_D1	다이어트 체중 운동 렌트 …
U2	K_D2	다이어트 음식 칼로리 …
U3	K_D3	캠핑 캠핑카 렌트 …
U4	K_D4	캠핑 캠핑장 텐트 …
U5	K_D5	캠핑카 텐트 렌트 …
U6	K_D6	맛집 연인 쇼핑 데이트코스 …

〈표 4〉 키워드 문서 분석을 통한 토픽 추출 가상 예

Topic ID	Topic Terms	Name	Member Keywords Documents
T1	다이어트, 체중, 운동, 칼로리, 식이요법	다이어트	K_D1, K_D2
T2	캠핑, 캠핑카, 캠핑장, 텐트, 렌트	캠핑	K_D1, K_D3, K_D4, K_D5
T3	맛집, 음식, 연인, 데이트코스, 데이트	데이트	K_D2, K_D6
T4	쇼핑, 쇼핑몰, 코트, 코디, 연예인	쇼핑	K_D6

위의 과정을 통해 형성된 키워드 문서들을 대상으로 토픽 분석을 수행함으로써 고객과 토픽 간 대응 매트릭스를 도출할 수 있다, <표 4>는 키워드 문서에 대한 토픽 분석을 통해 도출된 가상의 토픽을 보여주고 있으며, <그림 6>은 <표 4>를 활용하여 키워드 기반의 고객 클러스터링을 수행하는 과정을 네트워크의 형태로 보여주고 있다.



〈그림 6〉 검색 키워드 기반의 고객 클러스터링의 예

<그림 6>을 보면 Module 1과 Module 2의 예에서 상이한 그룹에 속해있던 “U3”, “U4”, “U5” 3명의 고객이 “캠핑”이라는 공통된 직접적 관심사를 가짐으로써 새로운 고객 그룹을 형성한 것을 알 수 있다. 즉, 인구통계학 정보가 아닌 관심사라는 새로운 기준으로 클러스터링을 수행함으로써 새로운 고객군을 식별할 수 있다.

4. 실험 및 결과

4.1 실험 데이터 소개

<그림 3>에 나타난 바와 같이, 본 연구의 실험을 위해 필요한 데이터는 크게 (1) Demography Information, (2) Web Transaction, 그리고 (3) News Articles로 요약될 수 있다. (1) Demography Information과 (2) Web Transaction의 경우, 국내 한 인터넷 사이트 순위 분석 전문 업체로부터 패널(고객) 5,000명의 인구통계학 정보 12개과 해당 패널의 2012년 7월 1일부터 2013년 6월 30일까지 1년 동안의 웹 사용 기록 143,293,592건에 대한 상세항목 18개를 제공 받았으며, 웹 사용기록의 경우 상세항목 중 고객 식별자와 URL만을 분석에 활용하였다. (3) News Articles의 경우, 웹 사용 기록 중 국내 최대 인터넷 뉴스 포털사이트의 접속 기록만을 추출하고, 해당 URL의 기사에 대한 크롤링을 수행하여 기사 394,303건을 분석하였다. 인터넷 뉴스 포털사이트의 경우, 주요 언론사의 기사를 취합하여 재공급하는 특성을 갖기 때문에 특정 매체의 시각에 편향되지 않은 기사를 제공한다는 장점을 갖는다. 또한 충분히 많은 기사 수와 접속 수를 기록하고 있다는 측면에서 본 실험의 대상 데이터로 적합한 특성을 갖고 있다고 할 수 있다.

4.2 Module 1 : 인구통계학 정보 기반의 고객 클러스터링

Module 1을 수행하기에 앞서, 인구통계학 정보를 가지고 있는 고객 5,000명 중 인터넷 포털사이트 뉴스 접속 기록과 키워드 검색 기록을 가지고 있는 고객 414명을 분석 대상으로 선정하였다. 선정된 고객 414명은 남성 65%와 여성 35%, 기

〈표 5〉 인구통계학 정보 기반의 고객 클러스터링 결과(일부)

Cluster_ID	USER_ID	CONTACT	GENDER	JOB	LOCATION	MARRY	PAY	SCHOOL	AGE	OS
2	1113	가정	여자	디자인직	서울	미혼	100~199만 원	대졸	30대	Windows NT 6.1
	2117	가정	여자	전산직	부산	미혼	없음	대졸	30대	Windows NT 6.1
	3610	가정	여자	자영업	서울	미혼	없음	대졸	30대	Windows NT 6.1
	4119	가정	여자	사무관리직	서울	미혼	100만 원 미만	학생(대학(원)생)	30대	Windows NT 6.1
	4459	가정	여자	영업/마케팅	광주	미혼	100만 원 미만	학생(대학(원)생)	30대	Windows NT 6.1
4	501	가정	남자	연구직	경상	기혼	500만 원 이상	대학원졸	40대	Windows NT 5.1
	536	직장	남자	사무관리직	서울	기혼	500만 원 이상	대학원졸	40대	Windows NT 5.1
	2836	가정	남자	금융	경기	기혼	500만 원 이상	대학원졸	40대	Windows NT 6.1
	3399	직장	남자	전산직	전라	기혼	300~399만 원	대학원졸	40대	Windows NT 5.1
	4714	직장	남자	자영업	서울	기혼	500만 원 이상	대학원졸	40대	Windows NT 5.1
20	1165	가정	남자	전산직	경기	기혼	400~499만 원	대졸	30대	Windows NT 5.1
	1268	직장	남자	사무관리직	경기	기혼	400~499만 원	대졸	40대	Windows NT 5.1
	1497	가정	남자	디자인직	경기	기혼	400~499만 원	대졸	30대	Windows NT 6.1
	2514	직장	여자	자영업	경기	기혼	400~499만 원	대졸	30대	Windows NT 6.1
	3459	가정	남자	사무관리직	경기	기혼	400~499만 원	대졸	40대	Windows NT 6.1

혼 49%와 미혼 51%로 구성되어 있었으며, 학력 수준은 대졸이 56%로 가장 많고 연령은 30대가 38%로 가장 많은 것으로 파악되었다. 선정된 고객들의 나이, 접속장소, 성별, 직업, 거주 지역, 결혼 유무, 사용 OS, 소득, 학력, 연령대의 12가지 인구통계학 정보를 활용하여 SAS Enterprise Miner 12.1에서 K-평균 군집화를 통해 전통적 방식의 인구통계학 정보 기반의 클러스터링을 수행하였다. 이 때, 각 클러스터링에 대해 최적 클러스터의 수를 분석 도구에 의해 추천 받는 대신, 반복 실험에 기반한 연구자의 판단을 통해 고객 그룹의 수를 20개씩으로 통일하였다. 이에 따른 Module 1의 인구통계학 정보 기반의 고객 클러스터링 결과가 <표 5>에 나타나있다.

4.3 Module 2 : 뉴스 접속 기록 기반의 고객 클러스터링

Module 2에서는 국내의 순위 분석 전문 업체로부터 제공 받은 웹 사용 기록 중 국내 최대 인터넷 포털사이트 뉴스 접속 기록만을 추출하고, 해당 URL의 기사에 대한 크롤링을 수행하여 총 394,303건의 뉴스 기사를 수집하였다. 토픽 분석은 Module 1에서 선정된 고객 414명의 인터넷 포털사이트 뉴스 접속 기록만을 추출하여 전체 뉴스 기사 중 3,162건에 대해 수행하였다. 즉, 전체 기사 중 고객 414명이 1년간 접속한 인터넷 포털사이트 뉴스 기사 3,162건에 대해 토픽 분석을 수행하여 총 50개의 토픽을 추출하였다(<표 6>).

〈표 6〉 뉴스 기사에 대한 토픽 분석 결과(일부)

Topic_ID	Name	Topic_ID	Name
1	안철수, 대선, 문재, 단일화, 문 후보	20	혐의, 검찰, 재판, 징역, 법원
2	보조금, 요금, 갤럭시, 가입자, 텔레콤	26	박시, 박시후, 사건, 시후, 혐의
4	애플, 갤럭시, 삼성전자, 아이폰, 스마트폰	30	주택, 가구, 아파트, 부동산, 대책
5	의원, 후보자, 장관, 대통령, 국회	31	건강, 병원, 치료, 환자, 피부
6	북한, 중국, 미사일, 한반도, 미국	33	정부, 대통령, 정책, 국민, 장관
7	경찰, 경찰서, 사건, 사고, 조사	40	한혜진, 기성용, 힐링캠프, 캠프, 열애
8	드라마, 시청률, 연기, 시청자, 작품	42	결혼, 조이뉴스, 게임, 열애, 서태지
10	코레일, 용산, 개발, 사업, 출자	45	애플, 특허, 소송, 삼성전자, 삼성

이 때, 토픽 분석의 품질을 향상시키기 위해 SAS Enterprise Miner 12.1의 Text Parsing 모듈과 서울대학교의 꼬꼬마 형태소 분석기를 활용하여 총 33,147개의 어휘로 구성된 불용어 사전을 구축한 후 분석에 활용하였다.

위의 과정을 통해 도출된 50개의 토픽에 근거하여 기사와 토픽 간 대응 매트릭스(<그림 7(b)>)를 구축한 후, 고객의 인터넷 포털사이트 뉴스 접속 기록을 분석하여 고객과 기사 간 대응 매트릭스(<그림 7(a)> 참조)를 도출하였다. 이렇게 도출된 두 가지 매트릭스를 병합하여 고객과 토픽 간 대응 매트릭스를 도출하였다(<그림 7(c)>). 이 때, 고객마다 접속한 뉴스 기사의 수가 모두 다르기 때문에 고객별로 표준화를 수행하여 표준화된 고객과 토픽 간 대응 매트릭스를 도출한 후(<그림 7(d)>), 뉴스 접속 기록 기반의 고객 클러스터링을 수행하

(a) Article

User

CUS_ID	news_2754	news_7192	news_7980	news_10840	news_12233
9	0	0	0	0	1
70	0	0	0	0	0
186	0	0	0	0	0
189	0	0	0	0	0
239	1	0	0	0	0
293	0	0	0	1	0
422	0	0	1	0	0

(b) Topic

Article

News_no	Topic_5	Topic_27	Topic_33	Topic_41	Topic_49	Topic_50
2754	0	1	0	0	0	1
7192	0	0	0	0	0	0
7980	1	0	0	1	0	0
10840	0	0	1	0	1	0
12233	0	1	0	0	0	0
12615	0	0	0	0	0	0
12623	0	0	0	0	0	0

(c) Topic

User

CUS_ID	Topic_5	Topic_27	Topic_33	Topic_41	Topic_50
9	0	1	0	0	1
70	0	0	0	0	1
186	0	0	0	0	0
189	0	0	0	0	0
239	0	1	0	0	1
293	12	7	11	11	2
422	1	0	0	1	0

(d) Topic

User

CUS_ID	Topic_5	Topic_27	Topic_33	Topic_41	Topic_50
9	0	0.0416667	0	0	0.0416667
70	0	0	0	0	0.0714286
186	0	0	0	0	0
189	0	0	0	0	0
239	0	0.5	0	0	0.5
293	0.0789474	0.0460526	0.0723684	0.0723684	0.0131579
422	0.2	0	0	0.2	0

〈그림 7〉 표준화된 고객과 토픽 간 대응 매트릭스 도출 과정(일부)

〈표 7〉 뉴스 접속 기록 기반의 고객 클러스터링 결과(일부)

Cluster ID	CUS_ID	Topics	Cluster ID	CUS_ID	Topics	Cluster ID	CUS_ID	Topics
4	501	안철수, 대선, 문재, 단일화, 문 후보 의원, 후보자, 장관, 대통령, 국회 정부, 대통령, 정책, 국민, 장관	8	190	박시, 박시후, 사건, 시후, 혐의 경찰, 경찰서, 사건, 사고, 조사 혐의, 검찰, 재판, 징역, 법원	18	225	회장, 사업, 그룹, 사장, 부회장 코레일, 용산, 개발, 사업, 출자 정부, 대통령, 정책, 국민, 장관
	840			586			249	
	981			629			415	
	1613			925			745	
	1702			1797			888	
	1781			2029			1113	
	1959			2087			1120	
	2193			2284			1189	
	2694			2500			1316	
	2892			2514			1893	
	3276			2840			2465	
	3313			3307			2639	
	3325			3475			3427	
	3565			3788			3482	
	4560			3809			3750	
	4645			4539			3763	

여 총 20개의 고객 그룹을 도출하였다(<표 7>).

4.4 Module 3 : 포털사이트 검색 키워드 기반의 고객 클러스터링

Module 3의 경우, Module 1에서 선정된 고객 414명의 인터넷 포털사이트 검색 기록을 분석하여 해당 고객이 직접 입력한 검색 키워드를 모두 추출한 후, 추출된 검색 키워드들을 고객별로 나누어 하나의 문서 형태로 통합하여 키워드 문서를 <표 8>과 같이 구축하였다. 이에 대해 토픽 분석을 수행하여 총 50개의 토픽을 추출하였고(<표 9>), 이에 근거하여 고객과 토픽 간 대응 매트릭스를 도출하였다. 이후 Module 2와 같은 방식으로 고객별 표준화를 수행하여 표준화된 고객과 토픽 간 대응 매트릭스를 도출한 후, 포털사이트 검색 키워드 기반의 고객 클러스터링을 수행하여 총 20개의 고객 그룹을 도출하였다(<표 10>).

〈표 8〉 고객별 키워드 문서(일부)

Cus_ID	Keywords Documents
9	힛더스팟 삼청점, 힛더스팟 삼청동, 힛더스팟 삼청 …
26	히어로즈워 접속불가, 히어로즈워 접속 …
42	휴대폰 가입 btv, 휴대폰 가격, 휴대폰 tv사은품 …
43	효율 캐싱, 효율 냉방, 효웅, 효용성, 효완, 효영 종영 소감 …
64	황정음 노출, 황정음 노민우 셀카, 황정음 김용준 결별 …
70	화려한, 화려하다, 화랑초등학교경쟁률, 화랑초등학교 …
91	홍석천 이이경, 홍석천 이은희, 홍석천 이상형 송승헌 …
103	홍대앞 맛집, 홍대알라또래, 홍대아가씨, 홍대스튜디오 …
106	홈텍스, 홈텍쇼ㅡ, 홈택스 종합소득세 신고방법 …
134	현대자동차, 현대자동, 현대자도??, 현대인재개발원 …
142	현대산업개발, 현대산업 아이파크, 현대사랑병원 …
160	현관에우, 현관방화문, 현관바닥타일, 현관바닥시트지 …
164	헬스효과, 헬스회원권 세무처리, 헬스트론 중고 …
174	헬로모바일, 헬로마미, 헬로라라, 헬로디바, 헬로네이쳐 …
186	헌터 레인부츠 오버진, 헌터 레인부츠 라임, 헌터 레인부츠 …

〈표 9〉 키워드 문서에 대한 토픽 분석 결과(일부)

Topic ID	Name	Topic ID	Name
1	+the, +you, +of, I, +to	20	복권, 연금, 연말정산, 역삼, 역대
2	김, 여자, 이혼, 결혼, 김연아	24	한국, 학교, 하하, 한강, 축구
4	남자, 여자, 쇼핑몰, 구제, 가을	25	창원, 남자, 착한, 송중기, 참치
6	엑셀, 한글, 엔, 무료, 숫자	28	기사, 필기, 사업자, 인터넷, 실기
8	아이패드, 아이유, 미니, 아이폰, 아이	30	테, 라, 아디다스, 남성, 티
9	올레, kt, 옥션, 올, 올림픽	33	코코, 코스트코, 코엑스, 코, 맨
12	부산, 부동산, 부산대, 부여, 아파트	37	드래곤, 플라이트, 드라마, 드레, 영화
13	토니, 토, 모리, 통영, 토토	44	재산, 인터넷, 장현, 재형저축, 아이폰
15	우리, 용인, 우리나라, 요즘, 용산	48	다크, 닥터, 천안, 단백질, 라이즈
18	코, 케, 케이, 팝스타, 슈퍼	50	커피, 컬, 커플, 컴퓨터, 투

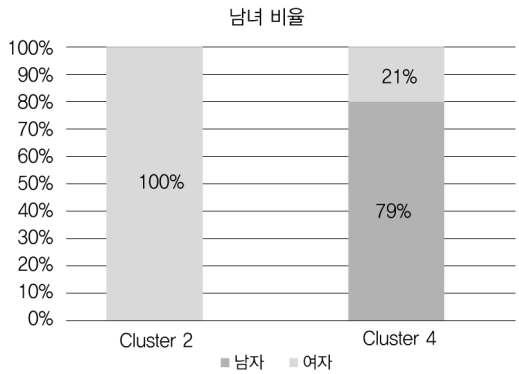
〈표 10〉 포털사이트 검색 키워드 기반의 고객 클러스터링 결과 (일부)

Cluster ID	CUS_ID	Topics	Cluster ID	CUS_ID	Topics	Cluster ID	CUS_ID	Topics
5	225	엑셀, 한글, 엔, 무료, 숫자 커피, 컬, 커플, 컴퓨터, 투 코, 케, 케이, 팝스타, 슈퍼	11	106	드래곤, 플라이트, 드라마, 드레, 영화 코코, 코스트코, 코엑스, 코, 맨 테, 라, 아디다스, 남성, 티	12	70	재산, 인터넷, 장현, 재형저축, 아이폰 기사, 필기, 사업자, 인터넷, 실기 복권, 연금, 연말정산, 역삼, 역대
	517			394			230	
	582			601			792	
	586			1613			847	
	1113			1797			926	
	1452			2085			1002	
	1614			2200			1074	
	1987			2267			1655	
	2072			2400			2389	
	2269			2828			2584	
	2842			3046			2762	
	3218			3325			2785	
	3276			3348			2833	
	3313			3397			3461	
	4459			3410			4072	

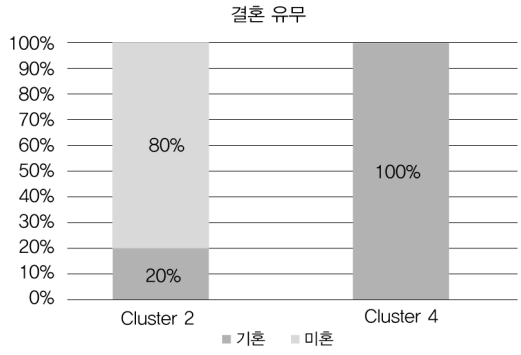
4.5 클러스터링 결과 비교

Module 1~Module 3의 과정을 통해 (1) 인구통계학 기반의 고객 클러스터링, (2) 뉴스 접속 기록 기반의 고객 클러스터링, (3) 포털사이트 검색 키워드 기반의 고객 클러스터링의 총 3가지 관점의 고객 클러스터링 결과가 도출되었다. 동일한 기준을 사용한 클러스터링의 경우 클러스터 내 동질성 및 클러스터간 이질성을 측정하여 성능 비교가 가능하지만, Module 1~Module 3은 서로 상이한 기준을 사용하여 클러스터링을 수행하였기 때문에 성능 비교가 용이하지 않다. 따라서 본 절에서는 클러스터링의 성능 비교가 아닌 결과의 상이성만을 확인하고자 한다. 또한 Module 3의 경우, 고객이 입력한 키워드를 그대로 문서로 정의하고 별도의 정제과정을 거치지 않았기 때문에 토픽분석 자체의 품질이 좋지 않게 나타났다. 따라서 본 절에서는 Module 1과 Module 2의 결과로 도출된 클러스터의 상이성을 확인하고자 한다.

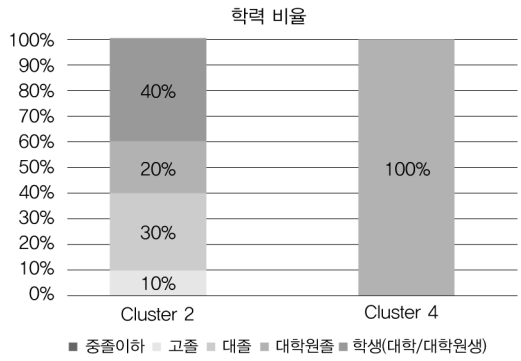
우선 Module 1의 경우, 대표적인 클러스터로 Cluster 2와 Cluster 4의 인구통계 비율을 비교해 보았다. <그림 8>과 <그림 9>를 보면 Cluster 2의 경우, 여성이 100%, 미혼자가 80%를 차지하는 것에 반해, Cluster 4의 경우에는 남성이 79%, 기혼자가 100%를 차지하는 것을 알 수 있다. 또한 학력 및 연령 비율 역시, Cluster 2의 경우에는 다양한 분포의 학력 비율을 가지면서 30대가 90%를 차지하는 것에 반해, Cluster 4의 경우, 대학원 졸업자가 100%, 30~40대가 각각 38%, 50%를 차지하는 것을 알 수 있다(<그림 10>, <그림 11>). 즉 두 클러스터는 인구통계 측면에서 서로 구별되는 특징을 가짐을 확인하였다.



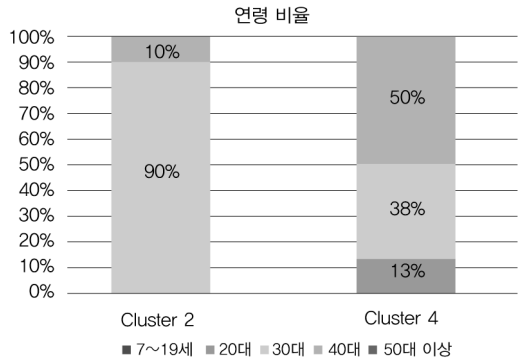
〈그림 8〉 인구통계학 정보 기반의 클러스터링 결과 중 Cluster 2와 Cluster 4의 남녀 비율



〈그림 9〉 인구통계학 정보 기반의 클러스터링 결과 중 Cluster 2와 Cluster 4의 결혼 유무 비율



〈그림 10〉 인구통계학 정보 기반의 클러스터링 결과 중 Cluster 2와 Cluster 4의 학력 비율



〈그림 11〉 인구통계학 정보 기반의 클러스터링 결과 중 Cluster 2와 Cluster 4의 연령 비율

Module 2의 경우에는 각각 “정치”, “스마트폰”, “연예인”이라는 서로 다른 토픽에 의해 그룹이 형성된 Cluster 4, Cluster 5, 그리고 Cluster 9(<표 11>)의 3개의 클러스터 간 인구통계 비율을 비교해 보았다.

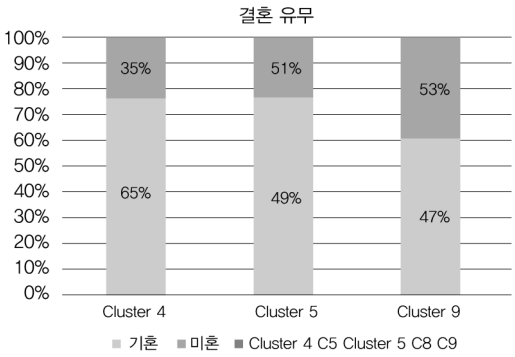
〈표 11〉 클러스터별 주요 관심사

Cluster ID	Topics	Interests
4	안철수, 대선, 문재, 단일화, 문 후보	정치
	의원, 후보자, 장관, 대통령, 국회	
	정부, 대통령, 정책, 국민, 장관	
5	애플, 갤럭시, 삼성전자, 아이폰, 스마트폰	스마트폰
	보조금, 요금, 갤럭시, 가입자, 텔레콤	
	애플, 특허, 소송, 삼성전자, 삼성	
9	한혜진, 기성용, 힐링캠프, 캠프, 열애	연예인
	결혼, 조이뉴스, 게임, 열애, 서태지	
	윤택, 빈소, 고인, 임윤택, 병원	

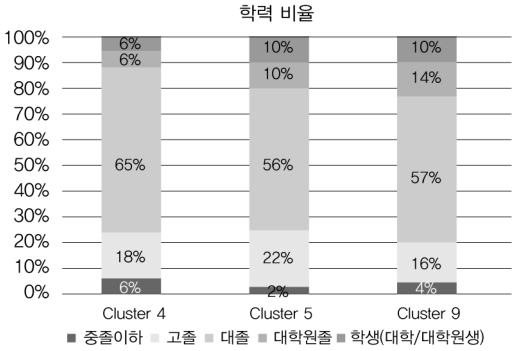
<그림 13>을 보면 Cluster 4에서는 기혼자가 65%인데 반해, Cluster 5와 Cluster 9에서는 기혼자가 각각 49%, 47%를 차지함으로써, Cluster 4가 조금 다른 특징을 보이고 있으나, <그림 12>와 <그림 14>를 보면 Cluster 4, Cluster 5, Cluster 9에서 남성이 각각 76%, 61%를 차지하고, 대졸자가 각각 65%, 56%, 57%를 차지함으로써 3개의 클러스터가 비슷한 인구통계 비율을 가지고 있는 것을 알 수 있다. 이처럼 뉴스 접속 기록 기반의 고객 클러스터링을 통해 도출된 클러

〈그림 12〉 뉴스 접속 기록 기반의 고객 클러스터링 결과 중 Cluster 4, Cluster 5, Cluster 9의 남녀 비율

스터들은 각각 서로 다른 관심을 갖는 고객들로 구성되어 있지만, 인구통계 관점에서는 서로 크게 구별되지 않음을 확인하였다.



〈그림 13〉 뉴스 접속 기록 기반의 고객 클러스터링 결과 중 Cluster 4, Cluster 5, Cluster 9의 결혼 유무 비율



〈그림 14〉 뉴스 접속 기록 기반의 고객 클러스터링 결과 중 Cluster 4, Cluster 5, Cluster 9의 학력 비율

5. 결 론

대부분의 기업들은 고객의 요구를 파악하여 맞춤형 제품과 서비스를 제공하기 위해, 특정 기준에 따라 고객을 여러 그룹으로 세분화하고 그 결과를 바탕으로 다양한 마케팅 전략을 수립하고 있다. 최근에는 많은 기업들이 고객의 인구통계학 정보뿐 아니라 고객의 구매 이력, 접속 패턴 등 다양한 정보를 활용하여 고객 세분화를 수행하고 있지만, 이들 정보의 대부분은 고객이 해당 사이트 내에서 보이는 특성을 나타낼 뿐 고객의 평소 성향 및 관심을 나타내기는 어렵다는 한계를 갖는다. 따라서 본 연구에서는 고객의 평소 인터넷 뉴스 조회 기록에 대한 텍스트 분석을 통해 고객의 실제 관심 분야를 식별하고, 이를 바탕으로 고객을 세분화할 수 있는 방안을 제시하였다. 또한 본 연구에서는 제안 방법론의 실제 적용 가능성을 평가하기 위해 인터넷 사용자 414명 및 이들이 조회한 뉴스 기사 약 3,000건에 대한 텍스트 분석을 수행하고, 그 결과 관심 기반 클러스터링을 통해 인구통계학 정보 기반 클러스터링과는 전혀 다른 고객군을 식별할 수 있음을 확인하였다.

본 연구의 기여는 더욱 정확한 고객 세분화 방안을 제시하는 것이 아니라 새로운 관점의 고객 세분화 방안을 제시한 것에서 찾을 수 있다. 즉 어떤 응용에서는 인구통계학 정보 또는 고객의 구매 이력 관점에서 동질성이 높은 고객군을 식별하는 것이 유용할 수 있지만, 또 다른 응용에서는 제안 방법론에 따라 관심이 유사한 고객군을 식별하는 것이 더욱 필요할 수 있다. 특히 제안 방법론은 인터넷 쇼핑몰의 추천시스템, 재방문 및 재구매 고객 예측 등 유사 관심 고객군의 식별이 필요한 응용 분야에 활용 가능성이 높을 것으로 판단되며, 이는 곧 본 연구의 실무적 기여로 인정받을 수 있다. 또한 학술적 측면에서 본 연구에서는 고객이 조회한 뉴스 기사 및 검색 기록만을 분석하여 고객의 관심을 식별하고 있으나, 향후 더욱 정교한 관심 기반 고객 세분화를 위해 인터넷 게시글, 소셜네트워크 게시글 등에 대해 제안 방법론을 적용한 후속 연구가 진행될 수 있을 것으로 기대한다.

일반적으로 클러스터링 방법론의 효과성 검증은 클러스터의 지름, 개체간 평균 거리 등 클러스터 내 동질성을 나타내는 척도와 근접 클러스터와의 중심 거리와 같이 클러스터간 이질성을 나타내는 척도를 사용하여 이루어진다. 하지만 이는 동일한

기준, 즉 동일한 변수 집합을 통해 수행된 여러 클러스터링 기법의 비교에 적용 가능하며, 본 연구와 같이 유사성을 파악하는 기준 자체를 새롭게 제시하는 방법론의 효과성 검증에는 적용하기 어려운 측면이 있다. 따라서 제안 방법론의 효과성 검증도 클러스터간 이질성 및 클러스터 내 동질성 측면이 아닌 실제 활용성 관점에서 이루어져야 하며, 이는 향후 연구에서 제안 방법론을 적용한 추천시스템의 성능 분석 및 고객의 재구매 예측 정확도 분석 등을 통해 이루어질 수 있다. 또한 향후 연구에서는 데이터 전처리, 분석 대상 기간 선정, 임계값 설정, 토픽 분석, 그리고 고객 클러스터링 등 분석의 모든 세부 과정에 대해 더욱 엄밀한 고찰이 이루어져야 하며, 이를 통해 제안 방법론을 적용한 결과의 품질 및 신뢰도를 더욱 향상시킬 수 있을 것으로 기대한다.

참 고 문 헌

[1] 고은주, 권준희, 윤선영, “라이프스타일에 따른 고객 세분화 및 e-CRM 전략제안”, *Journal of the Korean Society of Clithing and Textiles*, 제29권 제6호, 2005, pp. 847-858.

[2] 김세훈, “공공임대주택사업의 정책마케팅 적용 방안 연구 -정책고객의 세분화를 통한 마케팅 방향-”, *국가정책연구*, 제20권 제2호, 2006, pp. 57-91.

[3] 류 신, 김남규, “거시적 이슈 트래킹의 한계 극복을 위한 개인 관심 트래킹 방법론”, 2014 한국IT서비스학회 추계학술대회, 2014, pp. 534-543.

[4] 문권모, “고객 세분화, 다시 생각해야 한다”, LG주간경제, 제767호, 2004.

[5] 박광식, “항공사 고객가치 기반의 고객 세분화 사례연구”, *관광·레저연구*, 제26권 제1호, 2014, pp. 301-318.

[6] 박명호, 한장희, 김상우, 백운배, 인터넷 마케팅, 명경사, 2005.

[7] 송민영, “고객관계관리를 위한 고객 세분화 방법론 적용 : L사 사례”, 한국과학기술원 석사 학위논문, 2001.

[8] 안요찬, “고객 세분화를 통한 한방병원 고객관계관리 시스템 구축모형”, *한국산업정보학회 논문지*, 제15권 제5호, 2010, pp. 79-87.

[9] 오은녕, 이희상, “클러스터링 기법을 이용한 이동통신의 고객 세분화 연구”, 한국경영과학회 2002년 추계학술대회논문집, 2002, pp. 421-424.

[10] 윤종수, 윤종욱, “철강유통산업에서의 고객관계관리(CRM)를 위한 고객 세분화 모형”, *경영연구*, 제19권 제2호, 2004, pp. 144-164.

[11] 이건창, 정남호, “기계학습 기반의 웹 마이닝을 이용한 고객 세분화에 관한 연구”, *산업공학(IE interfaces)*, 제16권 제1호, 2003, pp. 54-62.

[12] 이두희, 통합적 인터넷 마케팅, 박영사, 2006.

[13] 임명수, 김남규, “기간별 이슈 매핑을 통한 이슈 생명주기 분석 방법론”, 2014 한국지능정보시스템학회 추계학술대회, 2014, pp. 291-299.

[14] 조홍규, “인공지능 방법을 이용한 신용평가 모형에 대한 개관”, 나이스채권평가 금융공학연구소, 2003.

[15] 정태수, 서의호, 황현석, 임승재, “이동통신회사에서의 Customer Value 측정을 통한 고객 세분화”, 한국경영과학회/대한산업공학회 2003년 춘계공동학술대회논문집, 2003, pp. 281-285.

[16] 진서훈, 안상욱, “신용카드업에서 데이터 마이닝의 활용-고객행동기반의 고객 세분화-”, 한국통계학회 2004년 학술발표논문집, 2004, pp. 171-174.

[17] 진서훈, “데이터 마이닝에 의한 고객 세분화

개발", *응용통계연구*, 제18권 제3호, 2005, pp. 555-565.

[18] 채경희, 김상철, "의사결정나무 기법을 활용한 백화점의 고객 세분화 사례연구", *유통과학연구*, 제8권 제1호, 2010, pp. 13-19.

[19] 한희섭, 황성훈, 황진수, "풀 서비스 레스토랑 산업에서 고객 세분화 방법을 이용한 고객 지향적 직원 서비스에 관한 연구", *호텔경영학연구*, 제22권 제6호, 2013, pp. 297-308.

[20] Albright, R., *Taming Text with the SVD*, SAS Institute Inc, 2006.

[21] Elmer, P. and Borowski, D., "An expert system approach to financial analysis, The case of S&L bankruptcy", *Financial Management*, 1988, pp. 66-75.

[22] Han, J. and Kamber, M., Data Mining : Concepts and Techniques, (3rd ed.), Morgan Kaufmann Publishers, 2011.

[23] Hearst, M. A., "Untangling Text Data Mining", in Proceedings of the 37th ACL, 1999.

[24] Mooney, R. J. and Bunescu, R., "Mining Knowledge from Text using Information Extraction", *ACM SIGKDD Explorations*, Vol. 7, No. 1, 2006, pp. 3-10.

[25] Provost, F. and Fawcett, T., "Data Science for Business", *O'Reilly*, 2013.

[26] Salton, G., Wong, A., and Yang, C. S., "A Vector Space Model for Automatic Indexing", *Communications of the ACM*, Vol. 18, No. 11, 1975, pp. 613-620.

[27] Sebastiani, F., "Machine Learning in Automated Text Categorization", *ACM Computing Surveys*, Vol. 34, No. 1, 2002, pp. 1-47.

[28] Sebastiani, F., "Classification of Text, Automatic", *The Encyclopedia of Language and Linguistics* 14, (2nd ed.), Elsevier Science Pub, 2006, pp. 457-462.

[29] Stanvrianou, A., Andritsos, P., and Nicoloyannis, N., "Overview and Semantic Issues of Text Mining", *ACM SIGMOD Record*, Vol. 36, No. 3, 2007, pp. 23-34.

[30] Witten, I. H., Text Mining, Practical Handbook of Internet Computing, CRC Press, 2004.

저자소개

현 윤 진

현재 국민대학교 비즈니스IT 전문대학원 박사과정에 재학 중이다. 국민대학교 비즈니스IT학부에서 학사 학위를 취득하고, 동 대학원에서 석사 학위를 취득하였다. 주요 관심분야는 텍스트 마이닝 및 데이터 마이닝이다.

김 남 규

현재 국민대학교 경영정보학부에서 부교수로 재직 중이다. 서울대학교 컴퓨터공학과에서 학사 학위를 취득하고, KAIST 테크노경영대학원에서 Database와 MIS를 전공하여 경영공학 석사 및 박사학위를 취득하였다. 한국정보기술응용학회 부회장, 한국경영정보학회 이사, 한국지능정보시스템학회 이사, 한국CRM학회 이사, 한국IT서비스학회지 편집위원, JITAM 편집위원, 한국인터넷정보학회논문지 편집위원을 역임하였으며, 한국생산성본부 TOPCIT 개발사업 자문위원으로 활동 중이다. 주요 관심분야는 텍스트 마이닝, 데이터 마이닝 및 데이터베이스 설계 등이다.

조 윤 호

현재 국민대학교 경영학부 빅데이터경영통계전공 교수로 재직 중이다. 서울대학교 계산통계학과를 졸업하고, KAIST 경영정보공학과에서 석사, KAIST 경영공학과에서 박사학위를 취득하였으며, LG전자(주)에서 6년간 주임연구원으로 재직하였다. 주 연구분야는 비즈니스애널리틱스, 빅데이터 마이닝, 추천시스템, 소셜네트워크분석, 고객관계관리 등이다.