토픽 분석을 활용한 웹 카테고리별 방문자 관심 이슈 식별 방안

최성이* · 김남규**

Identifying the Interests of Web Category Visitors Using Topic Analysis

Seongi Choi* · Namgyu Kim**

Abstract

With the advent of smart devices, users are able to connect to each other through the Internet without the constraints of time and space. Because the Internet has become increasingly important to users in their everyday lives, reliance on it has grown. As a result, the number of web sites constantly increases and the competition between these sites becomes more intense. Even those sites that operate successfully struggle to establish new strategies for customer retention and customer development in order to survive. Many companies use various customer information in order to establish marketing strategies based on customer group segmentation A method commonly used to determine the customer groups of individual sites is to infer customer characteristics based on the customers' demographic information. However, such information cannot sufficiently represent the real characteristics of customers. For example, users who have similar demographic characteristics could nonetheless have different interests and, therefore, different buying needs. Hence, in this study, customers' interests are first identified through an analysis of their Internet news inquiry records. This information is then integrated in order to identify each web category. The study then analyzes the possibilities for the practical use of the proposed methodology through its application to actual Internet news inquiry records and web site browsing histories.

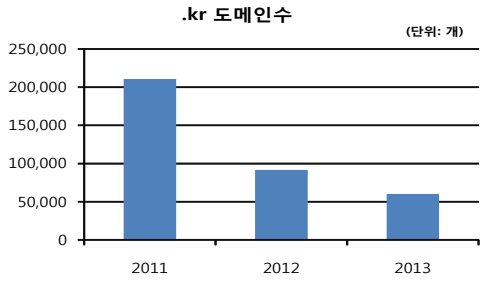
Keywords : Customer Segmentation, Site Categorization, Text Mining, Topic Analysis

논문접수일 : 2014년 11월 07일 논문수정일 : 2014년 12월 07일 논문게재확정일 : 2014년 12월 07일

* 국민대학교 석사과정, e-mail : csy0000@kookmin.ac.kr
** 교신저자, 국민대학교 경영정보학부 부교수, e-mail : ngkim@kookmin.ac.kr

1. 서 론

많은 사람들이 인터넷을 통해 정보를 습득하고 물건을 구매하는 등의 일상생활을 수행하고 있다. 특히 스마트 기기의 등장으로 인해 사용자들은 기존에 비해 시간과 공간의 제약 없이 인터넷에 접속할 수 있게 되었으며, 이로 인해 사용자들의 인터넷 의존도는 꾸준히 증가하고 있다. 그 결과 2013년 기준 전체 국민의 인터넷 평균 이용률은 82.1%로 나타났으며, 앞으로 더욱 높아질 것으로 전망되고 있다[KISA, 2014]. 이러한 수요 및 잠재 수요의 증가는 한때 인터넷 사용자들을 대상으로 수익을 창출하기 위한 인터넷 사이트가 우후죽순으로 생성되는 인터넷 골드러시 현상을 가져오기도 했다. 하지만 인터넷 시장이 성숙해짐에 따라 사이트 간 경쟁이 점차 심화되었으며, 경쟁 환경에 적응하지 못하는 사이트들은 도태되고 폐쇄되는 현상이 발생하고 있다(<그림 1> 참조).

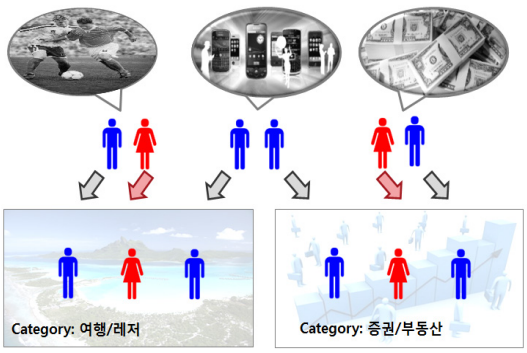


〈그림 1〉 최근 3년간 도메인수의 변화[KISA, 2014]

<그림 1>은 2011년부터 2013년까지 .kr 도메인 수가 꾸준히 감소하는 현상을 보이고 있다. 이러한 감소의 원인은 도메인 시장이 어느 정도 안정되었다는 점과 해당 도메인을 갖는 사이트들의 통합 및 폐쇄가 비일비재하게 발생한다는 점에서 찾을 수 있다. 특히 많은 수의 사이트들이 개설 후 한 달을 유지하지 못하고 주소가 폐쇄될 정도로 사이트 간 경쟁이 매우 치열하게 발생하고 있으며, 성공적으로 운영되고 있는 사이트들조차 지속적인 생존을 위해 고객 유지 및 신규 고객 발굴을 위한 전략 마련에 고심하고 있다.

경쟁력 강화를 위한 대부분의 전략은 기존 방문자들에 대한 특성을 분석하여 이들의 만족도를 향상시킬 수 있는 상품이나 서비스를 제안하는 방안을 모색하는 것을 목적으로 한다. 즉 사이트의 경쟁력 강화를 위한 대부분의 전략은 기존 방문 고객의 특성을 파악하는 것에서 출발한다고 해도 과언이 아니다. 실제로 많은 기업들은 기존의 고객 정보를 활용하여 고객군을 세분화 하고 이를 바탕으로 마케팅 전략을 수립하고 있다[편석환, 2005]. 개별 사이트 또는 매장에 대한 고객군의 파악이 용이하지 않을 경우 해당 사이트가 속하는 카테고리의 고객에 대한 정보를 바탕으로 개별 사이트 고객의 특성을 간접적으로 유추 하는 방법이 일반적으로 사용되고 있지만 이러한 분석 과정에서 사용되는 고객정보는 실제로는 고객의 인구통계적 정보에 국한되는 경우가 대부분이다. 예를 들면 “화장품 사이트를 주로 방문하는 고객은 20대 여성”과 같은 정보에 기반하여 타겟 마케팅을 실시하는 경우를 들 수 있다. 하지만 고객의 인구통계적 정보가 해당 고객의 특성을 충분히 대변한다고 볼 수는 없다. 화장품에 관심이 없는 20대 여성 고객도 많을 뿐 아니라, 20대 여성 고객은 화장품 이외에 쇼핑, 문화 등 다른 사이트에도 자주 방문할 수 있기 때문이다. 이는 인터넷 사이트의 카테고리별 특성을 방문고객의 인구통계적 정보에 기반하여 기술하는 방법의 한계를 보여주며, 이러한 한계는 <그림 2>를 통해 더욱 직관적으로 확인할 수 있다.

아래 <그림 2>는 남성 4명과 여성 2명의 사용자 접속 정보에 근거하여 여행/레저 카테고리와 증권/부동산 카테고리의 특성을 식별하는 간단한 예를 보여주고 있다.



〈그림 2〉 인구통계정보 기반 고객 세분화의 한계

인구통계적 관점에서 보면 각 카테고리에는 남성 2명과 여성 1명이 방문했으므로, 두 카테고리의 특성은 동일하다고 할 수 있다. 하지만 성별과 무관하게 고객들은 다양한 이슈에 관심을 갖고 있을 수 있으며, 이는 <그림 2>의 상단에 표시되어 있다. 고객의 관심 이슈 관점에서 보면 여행/레저 카테고리의 경우 스포츠에 관심을 갖는 고객이 많이 방문했고 증권/부동산 카테고리의 경우 경제에 관심을 갖는 고객이 많이 방문했으며, 스마트폰에 관심을 갖는 고객은 두 카테고리를 고르게 방문했음을 알 수 있다. 본 예는 고객의 인구통계적 정보뿐 아니라 고객의 관심 이슈에 대한 정보를 함께 고려함으로써 보다 정확한 고객 세분화를 수행할 수 있음을 의미한다.

이러한 필요에 기인하여 본 연구에서는 방문자 관심 이슈 분석을 통한 웹 카테고리 특성 식별 방안을 제시하고자 한다. 특히 방문자의 관심 이슈는 해당 카테고리에 속한 사이트의 방문 고객들이 평소 조회한 인터넷 뉴스 기사를 통해 획득하고자 하며, 관심 이슈는 최근 많은 연구가 이루어지고 있는 텍스트 마이닝(Text Mining)의 대표적 응용 중 하나인 토픽 분석(Topic Analysis)을 통해 파악하고자 한다. 본 논문의 이후 구성은 다음과 같다. 다음 장인 제 2장에서는 제안 기법과 관련된 기존의 연구를 소개하고, 제 3장에서는 본 논문에서 제안하는 방법론을 소개한다. 제 4장에서는 제안 방법론을 실제 사이트 및 뉴스 데이터에 대해 적용한 실험 결과를 소개하고, 마지막 장인 제 5장에서는 본 연구의 기여 및 한계를 요약한다.

2. 관련 연구

디지털 기술의 발달과 확산으로 인해, 과거에는 기록 또는 저장되지 않던 데이터들이 다량으로 저장되고 공유되고 있다. 이로 인해 데이터의 양 자체가 해결해야 할 문제의 일부로 여겨지는 빅데이터 분석[O'Reilly Radar Team, 2011]에 대한 수요 및 기술이 많은 주목을 받게 되었다. 구조화된 정형 데이터에 대한 분석을 통해 새로운 지식을 창출하는 위한 기존의 데이터 마이닝[Han et al., 2011]과는 달리, 텍스트 마이닝은 뉴스 기사, 웹 게시물, 소셜 미디어 상의 글 등의 텍스트 형태의 대규모의 비정형 데이터로부터 새로운 지식 및 유용한 패턴을 발견하기 위한 일련의 과정 및 기술을 의미한다. 구체적으로 텍스트 마이닝은 문서의 분류 및 군집화, 문서 요약, 정보 추출 등의 목적으로 활용되며[Tribula, 1999], 일반적으로 비정형 텍스트에 대한 전처리를 거쳐 이를 정형 데이터로 변환한 후 이후 분석 작업을 수행한다[Stanvrianou et al., 2007; Weiss et al., 2010]. 텍스트 데이터의 정형화를 위한 여러 기법들이 고안되어 왔으며, 특히 각 문서에 출현한 용어의 빈도를 요약하여 수치로 표시하는 벡터공간모델(Vector Space Model)[Albright, 2006; Salton et al., 1975]이 주로 사용된다.

텍스트 마이닝의 활용 분야는 텍스트 형태로 된 정보를 사용하는 모든 분야를 다룰 정도로 매우 다양하다. 구체적으로는 정보 추출, 정보 검색, 자연언어 처리, 텍스트 요약, 자동분류 등에서 사용되는 기법들을 결합해 특정 기사(Article)의 원문(Source)을 파악하기 위한 연구[Metzler et al.,

2005], 특정 범죄와 다른 범죄들 간의 유사성 측정을 통해 새로운 범죄를 발견하기 위한 연구[Fan et al., 2006], 텍스트 범주화(Categorization)를 통해 비구조적인 저장소를 구조화하기 위한 연구[Sebastiani, 2006], 검색 결과의 순위나 문서나 인식의 유사성을 구하는 연구[전채남, 서일원, 2013] 등이 있다. 특히 최근에는 사용자의 성향이 소셜 미디어를 통해 텍스트로 표현되고 저장됨에 따라, 다양한 소셜 미디어 데이터에 대한 텍스트 마이닝을 통해 기존의 데이터 분석에서는 찾을 수 없었던 새로운 유형의 지식을 찾기 위한 시도들이 활발하게 이루어지고 있다.

텍스트 마이닝의 다양한 응용 중 여러 분야에서 가시적인 성과를 내고 있는 대표적인 응용으로 토픽 분석을 들 수 있다. 토픽 분석은 문서를 분석의 최소 단위로 하여 수행되며, 이 때 문서는 문서, 제목, 요약, 본문, 댓글 등을 포함하는 넓은 개념으로 사용된다. 토픽 분석은 대량의 문서를 서로 유사한 문서들끼리 그룹화하고, 각 그룹을 설명할 수 있는 대표 키워드들로 해당 그룹을 기술하는 것을 목표로 수행되며, 하나의 문서가 여러 그룹에 동시에 속할 수 있다는 점에서 전통적인 군집분석과는 다른 특성을 갖는다.

토픽 분석을 활용하여 복잡한 문제를 해결하기 위한 시도가 최근 다양한 분야에서 활발하게 이루어지고 있다. 김지은 외[2014]에서는 토픽 분석을 통해 도출된 이슈의 수가 매우 방대한 경우 이슈에 대한 클러스터링을 통해 상위 이슈를 도출할 수 있으며, 이 때 단순히 이슈의 유사성이 아니라 사용자들의 이슈에 대한 관심 측면에서 상위 이슈를 도출하는 방안을 제시하였다. 현윤진 외[2013]에서는 다양한 국가 현안과 관련된 R&D 문서를 식별하고, 이를 패키지로 분류하여 보다 효율적인 검색이 이루어질 수 있는 방안을 제시하였으며, 홍진성 외[2014]에서는 토픽 분석을 활용하여 하나의 카테고리만 부여되어 있는 문서에 대해 자동으로 둘 이상의 카테고리를 부여할 수 있는 방안을 제시하였다.

토픽 분석에서 각 문서의 주요 용어의 선정은 용어의 출현 빈도수에 기반하여 이루어지며, 출현 빈도수는 그 목적에 따라 이진 모형(Binary Model), 삼진 모형(Three Value Model), 그리고 TF-IDF(Term Frequency-Inverse Document Frequency) 척도가 사용된다[Weiss et al., 2010]. 즉 각 문서는 "(문서 수)×(용어 수)"로 표현된 행렬의 각 셀에 문서에서 해당 용어가 나타난 빈도수를 기재함으로써 행렬로 정형화되며, SVD (Singular Value Decomposition) 등의 차원 축소 기법을 통해 저장되어 이후 분석에 활용된다[Albright, 2006].

3. 웹 카테고리별 사용자 관심 이슈 식별 방법론

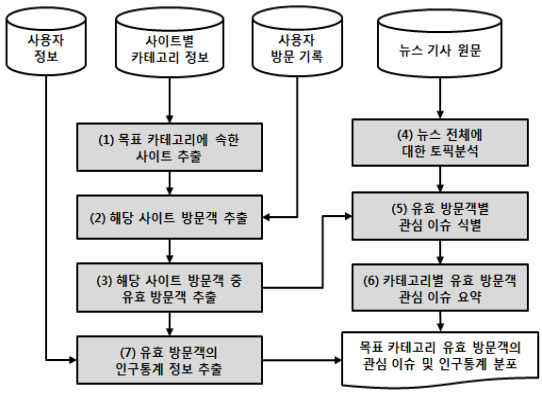
3.1 연구 개요

본 장에서는 뉴스 데이터에 대한 토픽 분석을 통해 사용자의 관심 이슈를 식별하고, 이를 해당 사용자의 웹 사이트 방문 기록과 연계하여 특정 웹 카테고리를 방문자의 관심 이슈 관점에서 특성화할 수 있는 방법론을 제시한다. 본 연구의 전체 개요는 <그림 3>과 같으며, 각 프로세스의 세부 과정에 대해서는 이후 절에서 상세히 소개한다.

<그림 3>에서 원통형으로 표시된 부분은 외부 데이터 소스를 의미하며 각각 사용자 정보, 사이트별 카테고리 정보, 사용자 방문기록, 그리고 뉴스 기사 원문을 나타낸다. 또한 직사각형으로 표시된 부분은 주요 프로세스를 나타내며, 우측 최하단의 도형은 최종 산출물을 나타낸다.

프로세스 (1)~(3)은 사이트별 카테고리 정보와 사용자 방문 기록으로부터 특정 카테고리에

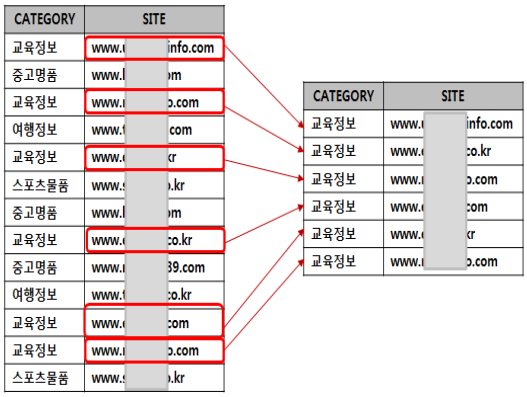
속한 사이트의 유효 방문객을 추출하는 과정이며 자세한 내용은 다음 절인 제3.2절에서 소개한다. 프로세스 (4)~프로세스 (5)는 뉴스 기사 원문에 대한 토픽 분석을 수행하고, 그 결과와 3.2절에서 추출한 유효 방문객 정보를 이용하여 유효 방문객별 관심 이슈를 식별하는 과정이다. 이 과정은 제 3.3절에서 다룬다. 유효 방문객별로 식별된 관심 이슈는 프로세스 (6)에서 카테고리별 방문 유효 방문객 관심 이슈를 요약하는 데에 사용되며, 사용자의 인구통계 정보와 3.2절의 유효 방문객 정보를 통합하여 유효 방문객의 인구통계 정보를 추출하는 과정은 프로세스 (7)에서 수행된다. 프로세스 (6)과 프로세스 (7)은 각각 제 3.4절과 제 3.5절에서 다룬다.



〈그림 3〉 연구 개요

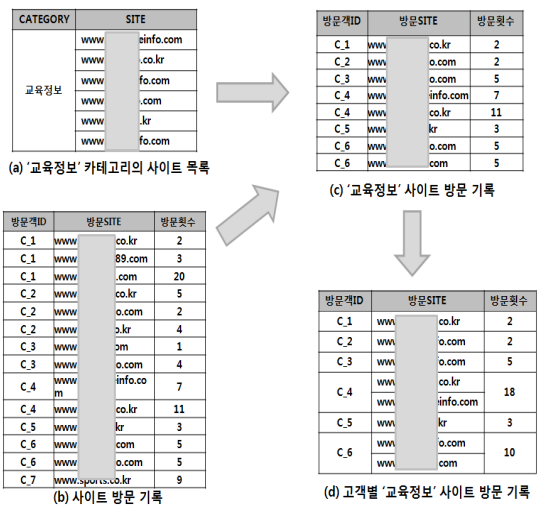
3.2 목표 카테고리의 유효 방문객 추출

본 절에서는 <그림 3>의 사이트별 카테고리 정보와 사용자 방문 기록으로부터 특정 카테고리에 속한 사이트의 유효 방문객을 추출하는 과정 (1)~(3)을 설명한다. 프로세스 (1)은 전체 사이트 정보 중 분석 대상 카테고리에 속한 사이트들만을 추출하는 과정이다. 예를 들어 <그림 4>는 '교육정보' 카테고리 사이트의 URL을 추출하는 예를 보여준다.



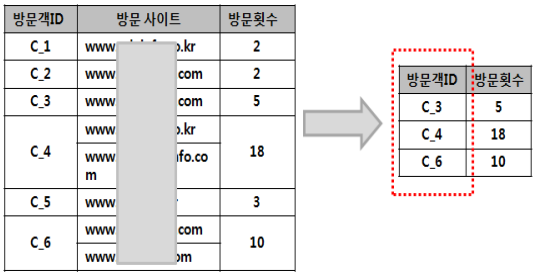
〈그림 4〉 목표 카테고리의 사이트 추출

프로세스 (2)는 프로세스 (1)에서 확인한 카테고리의 사이트를 방문한 사용자의 ID를 추출하는 과정을 나타낸다. <그림 5(a)>는 '교육정보' 카테고리의 사이트 목록을, <그림 5(b)>는 사용자의 사이트 방문 기록을 보여주고 있다. 이 두 테이블로부터 '교육정보' 사이트의 방문 기록을 추출한 결과가 <그림 5(c)>에 제시되어 있으며, 방문 기록을 각 방문객별로 집계한 결과가 <그림 5(d)>에 나타나있다. 즉 <그림 5(d)>는 '교육정보' 카테고리에 속한 사이트들의 방문 기록을 각 방문객별로 요약한 결과를 보여준다.



〈그림 5〉 방문객별 목표 카테고리 사이트 방문기록

본 연구에서 사용자의 사이트 방문은 사용자의 해당 사이트에 대한 관심을 나타내는 간접적인 척도로 활용된다. 하지만 사이트의 방문은 우연히 이루어질 수도 있고 일회성으로 발생할 수도 있기 때문에 모든 방문 기록이 사용자의 해당 사이트에 대한 관심을 의미하는 것으로 보기는 어렵다. 따라서 분석 결과의 신뢰성을 높이기 위해 유효 방문에 대한 임계값을 설정하고, 각 사용자가 특정 사이트에 대해 임계값 이상의 수만큼 방문한 경우만 유효 방문으로 인정한다. 이 과정은 하단의 <그림 6>에 소개되어 있으며, 유효 방문에 포함된 방문객의 ID만을 식별하여 추출한다.



방문객ID	방문 사이트	방문횟수
C_1	www .kr	2
C_2	www com	2
C_3	www com	5
C_4	www .kr	18
	www nfo.co m	
C_5	www	3
C_6	www com	10
	www m	

방문객ID	방문횟수
C_3	5
C_4	18
C_6	10

〈그림 6〉 목표 카테고리의 유효 방문객 추출

3.3 유효 방문객별 관심 이슈 식별

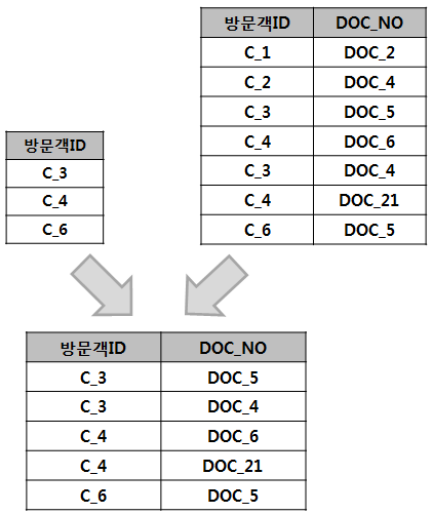
본 절에서는 프로세스 (4)~(5)를 통해 유효 방문객별 관심 이슈를 식별하는 과정을 설명한다. 프로세스 (4)는 수집한 뉴스 기사 전체에 대한 토픽 분석을 통해 주요 이슈를 추출하고 이슈별 주요 키워드를 요약한 뒤 각 이슈별 해당 문서를 제시한다. <그림 7>은 뉴스 기사에 대한 가상 토픽 분석을 통해 문서/이슈 대응 매트릭스를 도출한 결과를 보여준다. <그림 7>의 하단 매트릭스에서 특정 문서가 특정 이슈에 해당되는 경우 해당 셀의 값은 '1'로 나타난다. 예를 들어 DOC4는 이슈 T1과 이슈 T3에 대응된다. 토픽 분석은 이미 많은 연구를 통해 소개되었으므로, 자세한 분석 과정은 본 연구에서 다루지 않는다.

토픽ID	이슈 키워드	해당 문서
T1	기온, 지방, 날씨, 중부, 기상청	DOC2, DOC4, DOC6, DOC9, DOC11
T2	대출, 은행, 금융, 금리, 소득	DOC6, DOC32, DOC5
T3	여행, 지역, 마을, 사람, 버스	DOC26, DOC8, DOC4, DOC11, DOC9
T4	학생, 등교, 교육, 교사, 대학	DOC21, DOC5
T5	모델, 차량, 엔진, 자동차, 연비	DOC8, DOC45, DOC7

DOC_NO	T1	T2	T3	T4	T5
DOC_2	1	0	0	0	0
DOC_4	1	0	1	0	0
DOC_5	0	1	0	1	0
DOC_6	1	0	0	0	0
DOC_7	0	0	0	0	1
DOC_8	0	0	1	0	1
DOC_9	1	0	1	0	0
DOC_11	1	0	1	0	0
DOC_21	0	0	0	1	0
DOC_26	0	0	1	0	0

〈그림 7〉 뉴스 기사에 대한 가상 토픽 분석 예

다음으로 토픽 분석의 결과를 이용하여 유효 방문객의 관심 이슈를 식별하는 프로세스 (5)의 과정은 <그림 8>과 <그림 9>에 나타나있다. 우선 <그림 8>은 사용자의 뉴스 기사 방문 기록과 <그림 6>에서 도출한 유효 방문객의 ID를 이용하여 유효 방문객의 기사 조회 기록을 추출하는 과정을 보여준다.



방문객ID
C_3
C_4
C_6

방문객ID	DOC_NO
C_1	DOC_2
C_2	DOC_4
C_3	DOC_5
C_4	DOC_6
C_3	DOC_4
C_4	DOC_21
C_6	DOC_5

방문객ID	DOC_NO
C_3	DOC_5
C_3	DOC_4
C_4	DOC_6
C_4	DOC_21
C_6	DOC_5

〈그림 8〉 유효 방문객의 뉴스 조회 기록

다음으로 <그림 7>의 문서/이슈 대응 매트릭스와 <그림 8>의 유효 방문객의 뉴스 조회 기록을 사용하여 유효 방문객별 관심 이슈를 식별하게 되며, 이 과정은 <그림 9>를 통해 설명할 수 있다.

DOC_NO	T1	T2	T3	T4	T5
DOC_2	1	0	0	0	0
DOC_4	1	1	1	0	0
DOC_5	0	1	1	1	0
DOC_6	1	0	0	1	1
DOC_7	0	0	0	0	1
DOC_8	0	0	1	0	1
DOC_9	1	0	1	0	0
DOC_11	1	0	1	0	0
DOC_21	0	0	0	1	0
DOC_26	0	0	1	0	0

방문객ID	DOC_NO
C_3	DOC_5
C_3	DOC_4
C_4	DOC_6
C_4	DOC_21
C_4	DOC_5
C_6	DOC_6

방문객ID	DOC_NO	T1	T2	T3	T4	T5
C_3	DOC_5	0	1	1	1	0
	DOC_4	1	1	1	0	0
C_4	DOC_6	1	0	0	1	1
	DOC_21	0	0	0	1	0
	DOC_5	0	1	1	1	0
C_6	DOC_6	1	0	0	1	1

(a) 방문객의 문서별 이슈 대응도

방문객ID	T1	T2	T3	T4	T5
C_3	0	2	2	1	0
C_4	1	1	1	3	1
C_6	1	0	0	1	1

(b) 방문객별 이슈 대응도

〈그림 9〉 유효 방문객별 이슈 관심도

<그림 9(a)>는 문서/이슈 대응 매트릭스와 유효 방문객의 뉴스 조회 기록으로부터 방문객의 문서별 이슈 대응도를 생성하는 과정이다. 이 테이블에서 각 방문객마다 이슈별 대응도를 합산하여 방문객별 이슈 대응도를 산출할 수 있으며, 이는 <그림 9(b)>에 나타나있다.

3.4 목표 카테고리 유효 방문객 관심 이슈 요약

본 단계에서는 프로세스 (6)의 과정, 즉 목표 카테고리 유효 방문객의 관심 이슈를 요약하는 과정을 설명한다. <그림 9>의 최종 결과물은 목표 카테고리 유효 방문객 개인별 이슈 대응도를 나타낸다. 본 절에서는 개인별 이슈 대응도를 통합하여 목표 카테고리 유효 방문객 전체 차원의 관심 이슈를 요약할 수 있다. 이 과정은 <그림 10>에 요약되어 있다.

방문객ID	T1	T2	T3	T4	T5
C_3	1	2	2	1	0
C_4	1	1	1	3	1
C_6	1	0	0	1	1

(a) 방문객별 이슈 대응도

방문객ID	T1	T2	T3	T4	T5	sum
C_3	0.1666..	0.333..	0.333..	0.1666..	0	1
C_4	0.143..	0.143..	0.143..	0.429..	0.143..	1
C_6	0.333..	0	0	0.333..	0.333..	1
Sum	0.643	0.476	0.476	0.929	0.476	3

(b) 표준화된 방문객별 이슈 대응도의 총합

	T1	T2	T3	T4	T5	Sum
교육정보 카테고리 분포도	0.214 ...	0.158 ...	0.158 ...	0.309 ...	0.158 ...	1

(C) 목표 카테고리의 표준화된 관심 이슈

〈그림 10〉 교육정보 카테고리 분포도 평균

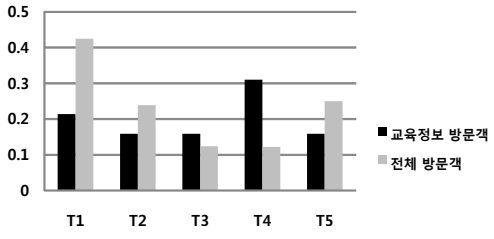
<그림 10(a)>는 방문객별 이슈 대응도를 나타내며 <그림 9(b)>와 동일하다. 여기서 방문객 C_3, C_4, 그리고 C_5가 갖는 관심 이슈 대응도의 총합은 각각 6, 7, 3으로 서로 상이하게 나타난다. 본 연구에서는 각 개인이 모든 이슈에 대해 갖는 관심의 총합은 동일하다고 가정하며, 이를 반영하기 위해 개인별 이슈 대응도의 표준화를 수행한다. 표준화는 <그림 10(a)>의 각 값을 각 개인이 갖는 관심의 총합으로 나누는 방식으로 수행하며, 그 결과가 <그림 10(b)> 상단에 제시되어 있다.

이렇게 표준화된 값에 대해 각 이슈 단위, 즉 각 열 단위의 총합을 구함으로써 목표 카테고리인 '교육정보' 카테고리에 속한 사이트 유효 방문객의 관심 이슈를 요약할 수 있으며, 그 결과가 <그림 10(b)>의 하단에 제시되어 있다. 마지막으로 <그림 10(b)> 하단 테이블의 각 값을 전체 값의 총합으로 나눔으로써 목표 카테고리를 구성하는 각 이슈의 비율을 표준화하여 나타낼 수 있다. 본 시나리오를 통해 소개된 가상 예의 경우 '교육정보' 카테고리에 속한 사이트 방문객은 이슈 T4인 '학생, 등교, 교육, 교사, 대학'에 가장 큰

관심을 보였으며, 이슈 T5인 '모델, 차량, 엔진, 자동차, 연비'에는 상대적으로 낮은 관심을 보였음을 알 수 있다. 또한 이슈 T1, T2, T3에는 비슷한 수준의 관심을 보였다

하지만 <그림 10(c)>에서 어떤 이슈가 높은 값을 가졌다고 해서, 해당 이슈가 목표 카테고리 방문객의 특징을 대표한다고 보기엔 무리가 따른다. 사이트 카테고리와 관계없이 많은 사용자로부터 큰 관심을 받는 이슈, 또는 그 반대의 이슈가 존재하기 때문이다. 따라서 방문객의 관심 이슈 분포에 근거하여 웹 카테고리의 특성을 파악하기 위해서는, 목표 카테고리 사이트 방문객에 대한 분석과 함께 전체 사이트 방문객에 대한 분석이 이루어져야 한다. 즉 전체 방문객의 관심 이슈 분포와 목표 카테고리 사이트 방문객의 관심 이슈 분포를 비교함으로써 해당 카테고리를 설명하는 이슈를 파악할 수 있으며, 이는 <그림 11>에 나타나있다. 예를 들면 <그림 10>에서 이슈 T1, T2, 그리고 T3의 값은 모두 동일하게 나타났지만, <그림 11>에서 전체 방문객의 관심에 대한 '교육정보' 방문객의 관심의 비율은 이슈 T3 > T2 > T1의 순으로 나타남을 알 수 있다.

<전체 방문객과 교육정보 방문객 비교 >



토픽ID	이슈 키워드
T1	기온, 지방, 날씨, 중부, 기상청
T2	대출, 은행, 금융, 금리, 소득
T3	여행, 지역, 마을, 사람, 버스
T4	학생, 등교, 교육, 교사, 대학
T5	모델, 차량, 엔진, 자동차, 연비

〈그림 11〉 목표 카테고리 유효 방문객 관심 이슈 분포

3.5 목표 카테고리 유효 방문객의 인구통계 분포

제안 방법론은 특정 웹 카테고리에 속한 사이트 방문객의 관심 이슈를 분석함으로써 해당 카테고리를 특성화하는 방안을 제시한다. 하지만 분석 결과를 해석하기 위해서는, 즉 목표 카테고리에서 특정 이슈의 관심도가 왜 높게 나타났는지를 설명하기 위해서는 추가 정보를 활용할 필요가 있다. 이를 위해 본 연구에서는 기존의 연구에서 일반적으로 사용되어 온 인구통계적 정보를 사용하고자 한다. 즉 제안 방법론은 토픽 분석을 통한 관심 이슈 정보 뿐 아니라 기존의 인구통계적 정보를 함께 활용함으로써, 목표 카테고리의 특성을 더욱 정교하게 파악할 수 있다. 이 과정은 본 장에서는 자세히 설명하지 않고, 다음 장에서 실험 결과만을 간략히 소개한다.

4. 실험 및 결과

4.1 실험 데이터 소개

본 장에서는 인터넷 순위 분석 전문 업체로부터 제공받은 사이트별 카테고리 정보, 사용자 방문 기록, 인구통계학 정보, 그리고 크롤링을 통해 수집한 인터넷 뉴스 기사 데이터를 활용하여 제안 방법론을 적용한 실험을 수행하고, 그 과정 및 분석 결과를 소개한다.

우선 사이트별 카테고리 정보에는 인터넷 사이트 150,295개를 1,251개의 카테고리로 분류한 정보가 포함되어 있다. 또한 사용자 방문 기록에는 패널 5,000명의 인터넷 방문 사이트, 방문 시각, 체류시간기록 약 1억 4천만 건이 포함되어 있다. 인구통계학 정보에는 위 패널 5,000명의 연령, 직업, 성별, 결혼유무, 학력 등의 정보가 포함되어 있다.

본 실험에서는 사용자의 관심 이슈 파악을 위해 인터넷 뉴스 기사를 활용한다. 즉 인터넷 뉴스에 대한 토픽 분석을 통해 주요 이슈를 도

출하고, 특정 이슈에 속한 기사를 다수 조회한 사용자는 해당 이슈에 관심이 높은 것으로 파악하는 것이다. 뉴스 기사는 다른 텍스트 문서에 비해 은어나 비속어의 사용이 적고, 표준 어휘와 정제된 표현을 사용하여 분석이 용이하다는 장점을 갖는다.

분석을 위해 2012년 7월부터 2013년 6월까지의 뉴스 기사 약 30만 건을 수집하였다. 이들 중 중 8개의 주제영역에서 각 3,000개씩, 총 24,000개의 기사를 추출하여 분석에 활용하였다.

4.2 목표 카테고리 유효 방문객 추출 결과

본 절에서는 목표 카테고리에 대한 유효 방문객을 추출한 결과를 소개한다. 본 실험에서는 특정 이슈에 대해 전체 패널에 비해 높은 관심을 보일 것으로 예상되는 카테고리인 '증권/투자정보'와 전체 패널과 유사한 관심 이슈 패턴을 보일 것으로 예상되는 카테고리인 'LOTTO정보'를 분석 대상 목표 카테고리로 선정하였다. 즉, 본 실험을 통해 '증권/투자정보', 'LOTTO정보', 그리고 전체 방문객의 관심 이슈 및 인구통계적 정보의 분포를 비교하고자 한다.

'증권/투자정보'와 'LOTTO정보' 카테고리에 속한 사이트의 수는 각각 370개와 164개로 나타났다. 이들 사이트에 접속한 방문객의 기록을 추출한 뒤, 각 카테고리별로 10개 이상의 사이트에 접속한 방문객만을 해당 카테고리의 유효 방문객으로 인정하여 추출하였다. 그 결과 각 카테고리별 유효 방문객의 수는 '증권/투자정보' 439명, 'LOTTO정보' 480명으로 파악되었다.

4.3 유효 방문객별 관심 이슈 식별 결과

본 단계에서는 SAS Enterprise Miner 12.1의 Text Miner 모듈을 사용하여 뉴스 기사 24,000건에 대한 토픽 분석을 수행한 결과를 소개한다. 전체 과정은 파싱(Parsing), 필터링(Filtering), 그리고 토픽 분석 순으로 이루어졌으며, 그 결과 <그림 12>와 같이 50개의 주요 이슈를 추출하였다.

<그림 12>는 주요 이슈 50개의 ID, 키워드, 그리고 해당 이슈의 문서 수를 보여주고 있다. 문서 수는 전체 기사 가운데 해당 이슈를 포함하는 기사의 수를 나타내며, 하나의 기사는 둘 이상의 이슈에 중복으로 포함될 수 있다.

토픽ID	이슈 키워드	문서수
1	갤럭시,삼성전자,스마트폰,제품	1994
2	대선,안철수,민주당,문재인,의원	2108
3	사람,생각,마음,자신	2845
4	경기,감독,리그,선수,박지성	1916
5	선수,감독,구단,시즌,롯데	1871
6	안타,타점,홈런,이닝,볼넷	1046
7	스타,드라마,방송,연예,리얼타임	2909
8	건강,환자,연구,질환,치료	2248
9	분기,시장,매출,실적,영업이익	2807
10	경찰,혐의,사건,범행,경찰서	2611
11	예금,특허,삼성전자,소송,삼성	1776
12	북한,개성공단,남북,회담,정부	1959
13	보조금,갤럭시,가입자,텔레콤	1816
14	경찰,의원,수사,혐의,의혹	2735
15	태풍,볼라,볼라벤,제주,강풍	920
16	류현진,다저스, 메이저리그,투수	1384
17	모델,자랑,엔진,자동차,연비	1800
18	당선인,인수위,대통령,박근혜	2464
19	요금,데이터,서비스,텔레콤,가입자	1811
20	올림픽,런던,금메달,대회,선수	1700
21	주택,아파트,부동산,가구,전세	1874
22	일본,중국,총리,센카쿠,정부	2587
23	사전,네티즌,커뮤니티,온라인	3018
24	대출,은행,금융,금리,소득	2361
25	제품,브랜드,매장,매출,고객	2928

토픽ID	이슈 키워드	문서수
26	코리아,게임,지디넷,애플,아이폰	1858
27	스타일,강남스타일,강남,곡	2205
28	통신사,공감,언론,이시간	2072
29	택시,정부,국회,의원,버스	3210
30	기온,지방,날씨,중부,기상청	883
31	증목,지수,코스피,외국인,주가	1628
32	정보,증권,스마트,경제증합	1075
33	경향,디지털,온라인,규직,저작권	1609
34	여행,지역,마을,사람,버스	2842
35	학생,학교,대학,교육,교사	2633
36	대변인,윤장,정와대,윤,대통령	2139
37	기업,회장,사업,그룹,회사	3170
38	나우뉴스,볼난,연구,통신원,볼난	2544
39	코레일,사업,용산,개발,출자	1805
40	앨범,전문 미디어,한국 최고	1849
41	발사,나로호,북한,미사일,우주	1732
42	정보,경우,인터넷,서비스,사용자	3134
43	경제,아시아,장 경제,총격,장	2030
44	게임,조이뉴스,메일,새로운시각	2319
45	미국,테러,보스턴,경찰,용의자	2819
46	계약,박지성,구단,맨유,선수	1821
47	리포트,아이폰,안드로이드,방송	2032
48	대통령,미국,공화당,대선	2837
49	김연아,피겨,쇼트,스케이팅,대회	1408
50	영상,페이스북,파일,복제,트위터	2837

〈그림 12〉 주요 토픽 50개의 이슈 키워드

<그림 12>을 통해 나타난 이슈와 문서간의 관계와 제 4.2절에서 도출한 유효 방문객의 기사 방문 기록을 연계하여 유효 방문객이 어떤 이슈의 기사를 조회했는지 파악할 수 있다.

전체 패널 5,000명 중 분석 대상 뉴스 기사 24,000건을 한 번이라도 방문한 패널은 2,723명으로 나타났다. 또한 '증권/투자' 카테고리의 유효 고객 439명 중 관심 이슈가 식별되는 고객, 분석 대상 뉴스를 한 건이라도 조회한 고객의 수는 286명으로 파악되었다. 마찬가지 이유로 본 실험에서는 'LOTTO' 카테고리의 유효 고객 각각 480명 중 301명 만의 관심 이슈를 파악할 수 있었다. 각

방문객의 이슈 대응도를 산출하고 이를 표준화한 결과가 <그림 13>에 나타나있다.

	방문객ID	T1	T2	T3	T4	T5	T6	T7	T8
1	1018	0	0	1	0	0	0	0	1
2	1023	0	0	2	0	0	0	1	1
3	1051	0	0	0	0	0	0	0	0
4	1054	0	1	0	1	1	0	0	0
5	106	0	0	1	0	0	0	4	0
6	1063	0	2	1	1	0	0	1	0
7	1067	2	11	5	5	6	4	2	4
8	1077	0	1	0	0	0	0	0	0
9	1078	1	0	0	0	0	0	0	0
10	1079	0	0	2	0	0	0	14	0
11	1105	60	5	11	1	2	0	4	7
12	1107	1	0	0	1	0	0	0	0
13	1124	0	0	0	0	0	0	0	0
14	1173	0	0	1	0	0	0	8	0
15	1216	2	1	0	0	0	0	0	0
16	1219	0	1	1	0	0	0	0	1
17	1229	1	4	4	5	5	2	1	2
18	1246	7	7	10	25	31	22	11	6
19	1260	2	1	2	0	0	0	0	4

(a) '증권/투자정보' 유효 방문객별 이슈 대응도

	방문객ID	T1	T2	T3	T4	T5	T6	T7	T8
1	1018	0	0	0.5	0	0	0	0	0.5
2	1023	0	0	0.083333	0	0	0	0.041667	0.041667
3	1051	0	0	0	0	0	0	0	0
4	1054	0	0.029412	0	0.029412	0.029412	0	0	0
5	106	0	0	0.047619	0	0	0	0.190476	0
6	1063	0	0.117647	0.058824	0.058824	0	0	0.058824	0
7	1067	0.011173	0.061453	0.027933	0.027933	0.03352	0.022346	0.011173	0.022346
8	1077	0	0.111111	0	0	0	0	0	0
9	1078	0.058824	0	0	0	0	0	0	0
10	1079	0	0	0.029412	0	0	0	0.205882	0
11	1105	0.07874	0.006562	0.014436	0.001312	0.002625	0	0.005249	0.009186
12	1107	0.083333	0	0	0.083333	0	0	0	0
13	1124	0	0	0	0	0	0	0	0
14	1173	0	0	0.015625	0	0	0	0.125	0
15	1216	0.083333	0.041667	0	0	0	0	0	0
16	1219	0	0.052632	0.052632	0	0	0	0	0.052632
17	1229	0.012821	0.051282	0.051282	0.064103	0.064103	0.025641	0.012821	0.025641
18	1246	0.016092	0.016092	0.022989	0.057471	0.071264	0.050575	0.025287	0.013793
19	1260	0.007547	0.003774	0.007547	0	0	0	0	0.015094

(b) '증권/투자정보' 유효 방문객별 표준화된 이슈 대응도

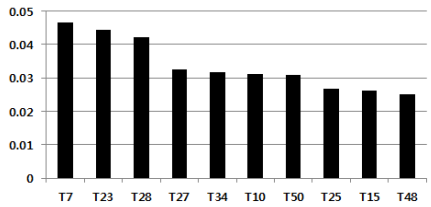
〈그림 13〉 '증권/투자정보' 카테고리 유효 방문객의 표준화된 이슈 대응도(일부)

<그림 13(a)>는 <그림 10(a)>에 대응되는 결과로, 각 방문객이 조회한 뉴스 기사와 해당 뉴스 기사가 갖는 이슈 대응도를 통해 산출되었다. 또한 각 방문객이 모든 이슈에 대해 갖는 관심의 총합을 '1'로 설정하여 표준화한 결과가 <그림 13(b)>에 나타나있으며, 이는 <그림 10(c)>에 대응되는 결과이다. 예를 들어 <그림 13(a)>에서 방문객 1018과 방문객 1023의 이슈 T8에 대한 관심도는 서로 동일하게 나타났지만, 표준화를 거친 이후인 <그림 13(b)>에서는 방문객 1018의 이슈 T8에 대한 관심도가 방문객 1023에 비해 높게 나타남을 알 수 있다. 'LOTTO정보' 카테고리의 경우도 이와 동일한 방식으로 유효 방문객의 표준화된 이슈 대응도를 산출하였으며, 자세한 과정에 대한 소개는 생략하도록 한다.

4.4 목표 카테고리 유효 방문객 관심 이슈 요약 결과

본 절에서는 제 4.3절에서 도출된 목표 카테고리 유효 방문객의 표준화된 이슈 대응도를 도식화하고, 카테고리별 관심 이슈 분포를 비교 분석한다. 우선 <그림 14>는 목표 카테고리와 무관한 전체 방문객의 관심 이슈 분포를 도식화한 그림이다. 각 이슈별로 전체 방문객이 갖는 관심도의 평균을 나타내고 있으며, 따라서 전체 50개의 이슈에 대한 관심도의 총합은 '1'로 유지된다. <그림 14>는 전체 50개의 이슈 중 평균 관심도가 높은 상위 10개의 이슈만을 보여주고 있다. 예를 들어 전체 50개 이슈 중 가장 관심도가 높게

< 전체 방문객의 관심 이슈 분포도 >

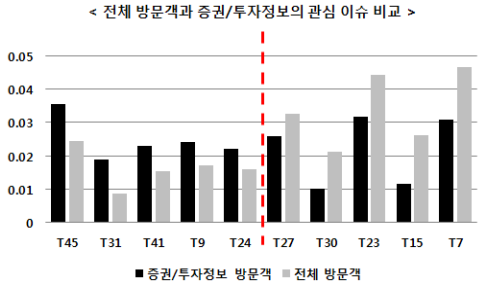


토픽ID	이슈 키워드
T7	스타,드라마,방송,연예,리얼타임
T23	사진,네티즌,커뮤니티,온라인
T28	통신사,언론,공감
T27	스타일,강남스타일,강남,가수,곡
T34	여행,지역,마을,사람,버스
T10	경찰,혐의,사건,범행,경찰서
T50	영상,페이스북,파일,복제,트위터
T25	제품,브랜드,매장,매출,고객
T15	태풍,볼라,볼라벤,제주,강풍
T48	대통령,룸니,미국,공화당,대선

〈그림 14〉 전체 방문객의 관심 이슈(상위 10개)

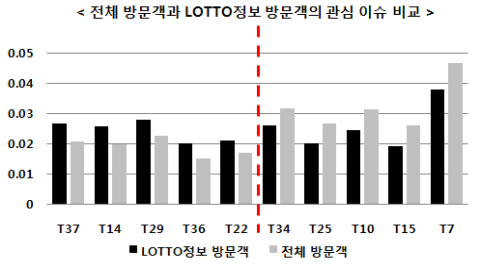
나타난 이슈는 '스타, 드라마, 방송, 연예'의 키워드를 갖는 이슈로, 해당 이슈에 대한 관심도의 평균은 0.047로 나타났다.

위 결과는 전체 방문객에 대한 관심 이슈를 보여주며, 특정 카테고리 유효 방문객의 관심 이슈와 비교하기 위한 목적으로 사용된다. 비교는 (i) 각 목표 카테고리별 방문객의 이슈별 관심도를 <그림 14>와 동일한 방식으로 산출하고, (ii) 각 이슈별로 목표 카테고리 유효 방문객의 관심도와 전체 방문객의 관심도의 차를 구한 뒤, (iii) 그 차가 크게 나타나는 이슈들을 요약하여 나타내는 방식으로 이루어진다. 특정 카테고리의 주요 관심 이슈는 해당 카테고리의 방문객들이 특별히 관심을 보이는 이슈 뿐 아니라, 모든 방문객들이 공통적으로 관심을 갖는 일반적인 이슈들도 포함될 수 있다. 따라서 특정 카테고리의 특성화를 위해서는 해당 방문객들의 주요 관심 이슈와 전체 방문객들의 관심 이슈간의 비교가 이루어져야 한다. <그림 15>는 전체 고객과 '증권/투자정보' 카테고리 방문객의 관심 이슈를 비교한 결과를 나타낸다. 또한 <그림 16>은 전체 고객과 'LOTTO정보' 카테고리 방문객의 관심 이슈를 비교한 결과이다.



토픽ID	이슈 키워드
T45	미국,테러,보스턴,경찰,용의자
T31	종목,지수,코스피,외국인,주가
T41	발사,나호로,북한,미사일,우주
T9	분기,시장,실적,매출,영업이익
T24	대출,은행,금융,금리,소득
T27	스타일,강남스타일,강남,가수,곡
T30	기온,지방,날씨,중부,기상청
T23	사진,네티즌,커뮤니티,온라인
T15	태풍,볼라,볼라벤,제주,강풍
T7	스타,드라마,방송,연예,리얼타임

〈그림 15〉 전체 방문객과 '증권/투자정보' 카테고리 방문객의 관심 이슈 비교



토픽ID	이슈 키워드
T37	기업,회장,사업,그룹,회사
T14	검찰,의원,혐의,수사,의원
T29	택시,정부,국회,의원,버스
T36	대변인,윤청,청와대,윤,대통령
T22	일본,중국,총리,센카쿠,정부
T34	여행,지역,마을,사람,버스
T25	제품,브랜드,매장,매출,고객
T10	경찰,혐의
T15	태풍,볼라,볼라벤,제주,강풍
T7	스타,드라마,방송,연예,리얼타임

〈그림 16〉 전체 방문객과 'LOTTO정보' 카테고리 방문객의 관심 이슈 비교

<그림 15>에서 '증권/투자정보' 카테고리 방문객이 전체 방문객에 비해 높은 관심도를 보이는 이슈는 이슈 T45, T31, T41, T9, 그리고 T24 순으로 나타났다. 반면 '증권/투자정보' 카테고리의 방문객이 전체 방문객에 비해 낮은 관심을 보인 이슈는 이슈 T7, T15, T23, T30, 그리고 T27로 나타났다. 전반적으로 해당 카테고리의 유효 방문객은 경제 관련 이슈 또는 주가에 영향을 미치는 대형 사건과 관련된 이슈에 대해 높은 관심을 갖는 반면, 연예 및 날씨 등의 이슈에는 상대적으로 낮은 관심을 보임을 알 수 있었다.

한편 <그림 16>에서 'LOTTO정보' 카테고리 방문객이 전체 방문객에 비해 높은 관심도를 보이는 이슈는 이슈 T37, T14, T29, T36, 그리고 T22 순으로 나타난 반면, 'LOTTO정보' 카테고

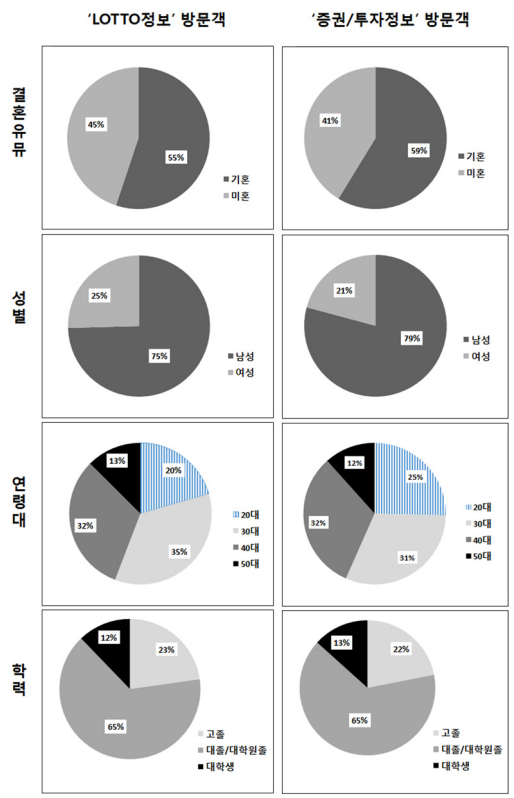
리의 방문객이 전체 방문객에 비해 낮은 관심을 보인 이슈는 이슈 T7, T15, T10, T25, 그리고 T34의 순으로 나타났다. 이 경우는 <그림 15>의 경우와 달리 해당 방문객이 더 많은 관심을 보이는 이슈들 간의 공통점이 명확히 나타나지 않았으며, 이는 'LOTTO정보' 카테고리의 방문객은 특정 이슈에 편향되는 특성을 보이지 않는 것으로 해석될 수 있다.

4.5 목표 카테고리 유효 방문객의 인구통계 분포 분석 결과

제안하는 방법론의 적용을 위한 주요 실험 결과는 제 4.2절~제 4.4절에 걸쳐 소개하였다. 본 절에서는 인구통계 정보에 기반하여 카테고리 방문객의 특성을 파악하는 기존의 방식을 적용해 봄으로써, 앞에서 제시된 제안 방법론의 적용 결과와 그 차이를 비교해 보고자 한다.

비교를 위해 '증권/투자' 카테고리 유효 방문객과 'LOTTO정보' 카테고리 유효 방문객의 두 그룹의 방문객 집단이 분석에 사용되었다. 즉 각 그룹에 속하는 방문객 집단의 인구통계 정보로부터 성별, 결혼유무, 연령, 그리고 학력 등의 분포를 요약하고, 이러한 분포의 차이를 통해 카테고리의 특성을 파악할 수 있는지 여부를 직관적으로 확인하고자 한다. 세 그룹의 성별, 결혼유무, 연령, 그리고 학력의 분포가 <그림 17>에 요약되어 있다.

<그림 17>의 결과는 두 카테고리 방문객의 인구통계적 분포의 차이를 도식화하여 보여주고 있다. 물론 엄밀한 비교는 통계적 유의성에 근거하여 이루어져야 하지만, 두 카테고리 방문객의 인구통계적 특성이 크게 이질적이지 않음은 그림을 통해 직관적으로 확인할 수 있다. 한편 <그림 15>와 <그림 16>을 통해 위의 두 카테고리의 방문객이 관심을 갖는 이슈는 서로 매우 다르게



〈그림 17〉 목표 카테고리별 유효 방문객의 인구통계 분포 비교

나타났다. 물론 각 집단간 인구통계 분포 차이의 유의성에 대한 통계적 검증이 이루어져야 하겠지만, <그림 17>은 본 논문에서 제안하는 방법론의 핵심이 아닌 단순 비교를 위한 결과이므로 이에 대한 엄밀한 분석은 본 연구에서 다루지 않았다. 본 실험을 통해서 방문객은 인구통계적 분포에 큰 이질성이 존재하지 않았지만, 카테고리 간 방문객이 실제로 관심을 갖는 이슈에서는 큰 차이가 존재할 수 있음을 확인하였다. 하지만 본 실험의 결과를 통해 인구통계 분포에 따른 고객 구분은 무의미하며, 관심 이슈에 기반한 고객 구분이 더욱 의미있다고 판단하는 것은 무리가 있다. 어떤 카테고리의 경우 인구통계 분포에 따른 고객 구분이 관심 이슈에 따른 구분보다 훨씬 유의한 결과를 보일 수도 있기 때문이다. 따라서 본

실험의 결과는 인구통계 분포에 따른 고객 구분과 관심 이슈에 기반한 고객 구분은 결과에 차이가 있을 수 있으며, 따라서 마케팅 전략 수립시 두 가지 기준에 대한 검토가 보완적으로 이루어져야 함을 암시하는 것으로 해석되는 것이 보다 바람직하다.

5. 결 론

많은 웹 사이트들은 해당 사이트 방문객의 정보를 활용하여 고객군을 세분화하고 이를 바탕으로 마케팅 전략을 수립하고 있다. 또한 개별 사이트의 고객 정보 분석이 용이하지 않을 경우 해당 사이트가 속하는 카테고리의 특성을 파악하고, 이를 바탕으로 개별 사이트 고객의 특성을 간접적으로 유추 하는 방법이 일반적으로 사용되고 있다. 하지만 카테고리의 특성 파악에 일반적으로 활용되는 정보는 방문객의 인구통계적 정보에 국한되는 경우가 대부분이며, 방문객의 인구통계적 정보가 해당 고객의 특성을 충분히 대변한다고 볼 수는 없다는 한계를 갖는다. 따라서 본 연구에서는 방문객의 인구통계적 정보 뿐 아니라 이들의 관심 이슈를 파악하여 웹 사이트 카테고리의 특성을 파악할 수 있는 방법론을 제시하였다. 즉 본 연구에서는 목표 카테고리에 속하는 사이트를 식별하고, 해당 카테고리에 속한 사이트 중 일정 수 이상의 사이트에 접속한 유효 방문객을 식별한 후, 이들의 인터넷 뉴스 접속 기록을 바탕으로 목표 카테고리의 특성을 유효 방문객의 관심 이슈 관점에서 기술하기 위한 세부 과정을 제시하였다. 제안 방법론은 특정 카테고리 방문객의 특성 파악을 위해 카테고리 내부 정보가 아닌 해당 방문객의 평소 웹 사용 기록을 사용했다는 점에서 기존 연구와의 차별성을 가지며, 특히 웹 사용 기록 중 뉴스 조회 기록에 대한 토픽 분석을 통해 사용자의 관심 분야를 요약하고 식별했다는 점에서 비정형 데이터에 대한 분석을 다루는 연구 분야에서의 기여가 인정될 수 있다.

제안 방법론의 실무 적용 가능성을 파악하기 위해 1,251개의 카테고리로 분류된 150,295개의 사이트에 대한 접속 기록, 그리고 인터넷 뉴스 기사 24,000건과 이 뉴스를 방문한 인터넷 사용자 2,723명에 대한 뉴스 조회 기록을 활용하여 실험을 수행하였다. 실험 결과 “증권/투자정보” 카테고리와 같이 방문객의 성향이 뚜렷한 경우 해당 카테고리 방문객의 관심 이슈는 전체 방문객의 관심 이슈와는 상이한 분포를 보였으며, “LOTTO 정보” 카테고리와 같이 방문객의 성향이 뚜렷하지 않은 경우는 해당 카테고리 방문객의 관심 이슈는 전체 방문객의 관심 이슈 분포와 큰 차이를 보이지 않았다. 한편 인구통계적 정보를 활용하여 카테고리 특성화를 시도한 비교 실험에서는 “LOTTO정보” 카테고리 방문객과 “증권/투자정보” 카테고리 방문객간에 성별, 결혼유무, 연령, 학력 면에서 카테고리간 별다른 차이가 나타나지 않았다. 따라서 제안 방법론이 기존의 인구통계적 정보 기반의 카테고리별 방문객 특성화의 한계를 극복하고, 목표 카테고리를 방문객이 실제로 관심을 갖는 분야의 관점에서 식별함으로써 보다 구체적이고 실현 가능한 마케팅 전략 수립에 활용될 수 있을 것으로 기대한다.

본 연구의 후속 연구에서 다루어져야 할 과제는 다음과 같다. 우선 이슈를 추출하기 위한 토픽 분석 과정에서 매우 정교한 전처리 작업이 요구된다. 특히 이슈 키워드로 도출된 용어 중에서도 일부 무의미한 용어가 발견되었는데, 이는 개체명 인식의 한계 및 한글 형태소 분석의 한계에 기인한 것으로 판단된다. 따라서 이슈 키워드의 품질 향상을 위해서는 양질의 용어사전 및 불용어사전의 구축이 반드시 선행되어야 할 것이다. 또한 본 연구에서는 제안 방법론의 실무 적용 가

능성을 가늠하기 위해 두 개의 목표 카테고리에 대한 분석만을 수행하였다. 하지만 제안 방법론을 통해 실제로 카테고리를 특성화하고 이를 통해 마케팅 전략을 수립하기 위해서는, 충분히 많은 수의 카테고리에 대한 추가 실험 및 이들 실험에 대한 결과 분석이 이루어질 필요가 있다.

참 고 문 헌

[1] 김지은, 김남규, 조윤호, "다계층 이원 네트워크를 활용한 사용자 관점의 이슈 클러스터링", *지능정보연구*, 제20권 제2호, 2014, pp. 93-107.

[2] 전채남, 서일원, "빅데이터 분석의 기술마케팅 활용에 관한 연구 : 잠재 수요기업 발굴을 중심으로", *한국전략마케팅학회지*, 제21권 제2호, 2013, pp. 184-187.

[3] 편석환, "인터넷 사이트 이용자 특성에 따른 광고영상 비교를 통한 광고효과 연구", *한국컨텐츠논문지*, 제5권 제2호, 2005, pp. 69-77.

[4] 현윤진, 한희준, 최희석, 박준형, 이규하, 곽기영, 김남규, "텍스트 분석을 활용한 국가 현안 대응 R&D 정보 패키징 방법론", *Journal of Information Technology Applications and Management*, 제20권 제3호, 2013, pp. 231-257.

[5] 홍진성, 김남규, 이상원, "단일 카테고리 문서의 다중 카테고리 자동확장 방법론", *지능정보연구*, 제20권 제3호, 2014, pp. 77-92.

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

[7] Fan, W., Wallace, L., Rich, S., and Zhang, Z., "Tapping the Power of Text Mining", *Communication of the ACM*, Vol. 49, No. 9, 2006, pp. 76-82.

[8] Han, J., Kamber, M., and Pei, J., *Data Mining : Concepts and Techniques*, 3rd Edition, Morgan Kaufmann Publishers, 2011.

[9] KISA, ".kr도메인통계", 2014 (available at-http://isis.kisa.or.kr/).

[10] O'Reilly Radar Team, *Big Data Now : Current Perspectives from O'Reilly Radar*, O'Reilly, 2011.

[11] Sebastiani, F., "Classification of Text, Atomatic", *The Encyclopedia of language and Lingustics* 14, 2nd edition, Elsevier Science Pub, 2006.

[12] 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.

[13] 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.

[14] Tribula, W. J., "Textming", *Annual Review of information science and Technology*, Vol. 34, 1999, pp. 184-187.

[15] Weiss, S. M., Indurkhya, N., and Zhang, T., *Fundamentals of Predictive Text Mining*, Springer, 2010.

저자소개

최 성 아

현재 국민대학교 비즈니스IT 전문대학원에서 비즈니스IT를 전공하고 있다. 원광대학교 정보전자상거래학사 학위를 취득하였으며, 주요 관심분야는 텍스트 마이닝, 토픽 모델링 및 데이터베이스 등이다.

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 시맨틱 데이터 관리 등이다.