KIPS Trans. Softw. and Data Eng.
Vol.10, No.6 pp.223~234
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2021.10.6.223

A Study on Search Query Topics and Types using Topic Modeling and Principal Components Analysis

Hyun-Ah Kang† · Heui-Seok Lim††

ABSTRACT

Recent advances in the 4th Industrial Revolution have accelerated the change of the shopping behavior from offline to online. Search queries show customers' information needs most intensively in online shopping. However, there are not many search query research in the field of search, and most of the prior research in the field of search query research has been studied on a limited topic and data-based basis based on researchers' qualitative judgment. To this end, this study defines the type of search query with data-based quantitative methodology by applying machine learning to search research query field to define the 15 topics of search query by conducting topic modeling based on search query and clicked document information. Furthermore, we present a new classification system of new search query types representing searching behavior characteristics by extracting key variables through principal component analysis and analyzing. The results of this study are expected to contribute to the establishment of effective search services and the development of search systems.

Keywords : Search Query Types, Text Mining, Topic Modeling, PCA, Log Analysis

토픽모델링 및 주성분 분석 기반 검색 질의 유형 분류 연구

강 현 아† · 임 희 석††

요 약

4차 산업 혁명 시대의 도래에 따라 쇼핑의 행태는 더욱 빠르게 오프라인에서 온라인으로 이동하고 있다. 온라인 쇼핑에서 고객의 정보요구를 가장 집약적으로 보여주는 것이 바로 검색 질의이다. 하지만 검색 분야에서도 검색 질의 관련 연구 사례는 많지 않으며 대부분의 검색 질의 연구 분야 선행 연구들은 연구자의 정성적인 판단에 근거하여 제한적인 주제와 데이터 기반으로 연구되어 왔다. 이에 본 연구는 검색 질의 연구 분야에 기계학습을 적용하여 검색 질의와 검색 이후 이용자가 조회한 문서명 로그를 기반으로 토픽모델링 수행 후 검색 질의 주제를 정의함으로써 데이터 기반의 정량적 방법론으로 15개의 검색 질의 주제 유형을 정의하였다. 또한 기존 검색어 자체만을 보고 판단하던 주제 유형에서 나아가 검색 행동특성을 반영한 유형을 정의하기 위하여 주성분 분석을 통해 주요 변수를 추출 후 각 주제별 검색 행동특성을 분석함으로써 검색 탐색 활성도, 상품 관여도에 따른 4가지의 새로운 검색 질의 유형 분류체계를 제시하였다. 본 연구결과는 효과적인 검색서비스 구축 및 검색 시스템 개발에 기여할 것으로 기대된다.

키워드 : 검색 질의 유형, 텍스트마이닝, 토픽모델링, PCA, 로그 분석

1. 서 론

국내 온라인커머스 시장은 매년 두 자릿수의 높은 성장률을 보이며 성장하고 있으며 통계청 자료에 따르면 2019년도는 매출액 기준 유통시장의 점유율 28.2%를 차지하여 오프라인 대형마트의 아성을 무너뜨리고 있다. 특히 2020년 발생한 COVID-19 바이러스로 인해 모든 일상생활이 비대면화 되면서 비대면으로 쇼핑이 가능한 온라인커머스의 성장이 가속화되고 있다. 이처럼 온라인커머스의 급성장에 따른 이와 관련된 다양한 연구들이 진행되었다.

권혁인 외 3명의 연구[1]에 따르면 e-커머스의 산업 생태계의 활성화 요인을 가중치 내림차순으로 '검색서비스 개발(0.0970)', '추천서비스 개발(0.0805)', '소비자 니즈 분석(0.0534)', '고객 소비 패턴 분석(0.0505)', '타 플랫폼 연계 서비스 개발(0.0450)'로 선정하였다. 해당 연구에서 언급된 요인 중 검색엔진, 검색 시스템 또는 추천서비스 연구 등은 이미 학계, 산업에서 연구가 활발한 분야인 반면 '소비자 니즈 분석'의 관한 연구는 선행 연구가 많지 않은 실정이다. 특히 온라인커머스에서의 소비자 니즈의 구체적인 발현은 검색 질의라고 할 수 있는데 국/내외 대부분 연구들이 제한적인 데이터를 대상으로 연구자의 정성적인 판단에 근거하여 검색 질의 유형을 분류하는 방법으로 연구되어 왔다.

※ 이 논문은 2020년 한국정보처리학회 추계학술발표대회의 우수논문으로 "토픽모델링과 주성분 분석을 활용한 온라인 쇼핑 검색 질의 유형 분류"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 고려대학교 빅데이터융합학과 석사과정
†† 정 회 원 : 고려대학교 컴퓨터학과 교수
Manuscript Received : February 24, 2021
Accepted : March 26, 2021
* Corresponding Author : Heui-Seok Lim(limhseok@korea.ac.kr)

본 연구는 온라인커머스에서 고객의 정보요구(Information needs)를 가장 집약적으로 나타내는 검색 질의에 대해 빅데이터 기반 기계학습을 활용하여 정량적인 방법으로 유형과 특성을 분석하고자 한다. 분석 데이터는 2019년 월평균 세션 수 2.6억 규모의 국내 온라인커머스 사이트에서 2019년 1년간 발생한 빅데이터 검색 로그를 활용한다. 연구는 크게 2단계로 구성된다.

첫 번째 단계는 토픽모델링을 통한 검색어 주제 유형 분류이다. 검색 질의 주제 유형 분류는 검색 질의(query)와 검색 후 조회 문서명의(상품명) 비정형 텍스트데이터를 수집하고 이를 문서명과 문서 내용의 관계로 간주하여 텍스트에서 자동으로 주제를 추출해주는 토픽모델링을 수행한다. 토픽모델링 기법으로 검색 질의를 분류하는 아이디어는 검색 질의를 문서명으로 간주하고 해당 질의 이후 클릭 된 상품명을 문서로 간주하였을 때 클릭 되는 상품명이 유사하다면 질의 간 유사도 또한 높을 것이라는 점에 착안해 구상되었다. 이를 통해 검색 질의 자체뿐만 아니라 검색 이후 조회된 문서까지 고려하여 소비자의 의도를 명확하게 담아 질의 유형을 정의할 수 있으며 기계학습 모델을 사용하여 자동으로 분류하기 때문에 빅데이터에 기반하여 정량적인 방법론으로 연구할 수 있다는 데 연구의 차별점이 있다.

두 번째 단계는 토픽모델링 기반으로 얻어진 검색 질의 주제 유형에 대해 주제별 검색 행동특성을 분석한다. 검색 행태 관련 변수를 집계하고 이에 대해 주성분 분석(PCA)을 수행하여 검색 행동특성별로 검색 질의를 분류할 수 있는 체계를 제시하고자 한다. 검색 이후 검색결과 세션에서 발생하는 검색 행동 로그를 수집하여 검색결과 클릭률, 평균 클릭문서 위치, 검색 후 구매시도율 등의 검색 행동을 대표하는 변수 총 12개를 집계한다. 집계된 12개 변수 대상으로 주성분 분석을 진행하여 제 1, 2 주성분을 도출하고 주성분의 특성을 명명한다. 이후 두 주성분을 x축, y축으로 한 직교좌표평면에 15개의 검색 주제 유형을 투사하여 총 4개의 행동특성별 유형을 정의하고자 한다. 본 연구는 검색 질의 자체가 갖는 정보요구뿐만 아니라 검색 이후 검색 세션 내 검색결과와 상호작용 하는 로그를 기반으로 검색 질의별 검색행태에 대한 체계적인 분석이 가능하며 이를 통해 선행 연구에서 다루지 않았던 새로운 검색 질의 분류체계를 제시하고자 한다.

제2장에서는 관련 연구 및 이론에 대해서 다루고 제3장에서는 연구 프레임 및 방법론을 기술한다. 제4장에서는 연구결과에 대해 분석하고 제5장에서는 결론 및 향후 연구 과제에 대해 기술한다.

2. 관련 연구

웹 검색 분야에서 빅데이터 수준의 트랜잭션(transaction) 로그를 활용하여 검색어를 분석한 연구 관련하여 Silverstein et al.(1999)이 1998년 8월 2일부터 6주간의 알타비스타 이용자들이 남긴 약 3억 개 수준의 이용자 세션과 약 10억 개의 질의를 분석하였다[2]. 해당 연구는 지금까지 트랜잭션 로그를 활용한 연구 중 가장 방대한 데이터를 기반으로 진행된 연구였고 세션 정의 등과 같은 로그 분석 방법론을 제시하였다는데 의의가 있다. Spink et al.(2001)은 1997년 9월 16일 Excite 웹페이지 이용자들이 남긴 약 100만 개의 질의 대상으로 2,414개를 무작위로 추출 후 이를 11개의 범주로 분류하는 체계를 도출하였다.[3] 11개의 범주는 엔터테인먼트, 성/성인, 상업/여행/고용/경제, 컴퓨터/인터넷, 건강/과학, 사람/장소/사물, 사회/문화/인종/종교, 교육/인문학, 예술, 정부, 불분명으로 구성되어 있다. 이후 해당 연구를 이용하여 시계열 변화 추이 연구를 진행하였다. Spink et al.(2002)은 1997년부터 2001년까지 2년에 한 번씩 무작위로 하루를 선정하여 20만 이용자로부터 얻은 약 100만 개의 검색 질의 중 무작위로 2,500개의 추출하여 주제를 분류하였다[4]. 그 결과 검색 질의 주제가 엔터테인먼트, 성 관련 주제로부터 전자 상거래 관련 주제로 변화하였으나, 전반적인 검색행태는 변하지 않았음을 보고하였다. Jansen, Spink, Pedersen(2005)은 2002년 9월 8일 알바스타에서 생성된 약 100만여 개의 질의로부터 약 2,600여 개를 무작위로 추출한 뒤 주제를 분류하고 이를 Silverstein et al.(1999)의 연구결과와 비교하였다[5]. 이들은 2002년의 질의 결과가 1998년보다 더 다양해지고 광범위해졌으며 성과 관련된 질의들이 감소하고 엔터테인먼트 관련 질의가 증가하였다고 설명하였다. 하지만 Silverstein의 연구에서는 질의 주제를 분류하지 않고 검색횟수가 높은 질의들을 대상으로 분석하였기 때문에 두 연구자료를 비교하는 것은 적절하지 않은 것으로 보인다. Rose와 Wolfram(2000)은 Excite 웹페이지의 검색엔진에 생성된 2만 개의 질의를 기반으로 질의 분석을 수행하였다[6]. 이들은 2개 이상의 검색어로 구성된 질의들을 추출하고, 이들 중 가장 많이 함께 출현하는 1,054개의 질의 쌍들의 유형을 카테고리화 하였다. 질의의 빈도와 동시 출현빈도를 계층적 클러스터링 분석 기법을 통해 군집화하는 귀납적인 방법론으로 조사하였다. 그 결과 주제들을 30개 유형으로 분류하였는데 성, 집단, 장소, 그림, 기관, 교육, 무료, 무역, 컴퓨팅, 인물, 웹/네트워크, 직업/경영, 멀티미디어, 음악, 참고, 커뮤니케이션, 뉴스, 출판물, 정부/법, 게임, 스포츠, 여행, TV/여행, 시각예술, 건강/의학, 역사, 이야기, 동물, 과학, 게임/복권 등으로 구성되어 있다. 이들은 질의 자체만 보고 주제를 도출하거나 질의만으로 알 수 없는 경우는 실제 사이트에 해당 질의를 입력하여 출력된 결과물을 바탕으로 주제를 분류하였다.

국내에서는 박소연, 이준호, 김지승(2005)이 2003년 7월부터 2004년 6월까지 1년간 네이버에서 발생한 18,200개의 질의 로그와 검색 이후 검색결과에서 이용자가 조회한 문서 등의 로그를 바탕으로 질의 주제를 분류하였다[7]. 질의의 형태별 분류로는 사이트 검색, 내용 검색 2개의 카테고리로 나누었고 사이트 검색의 질의 수가 내용 검색보다 더 많다고 기

술하였다. 질의 주제별 분류는 총 16개로 건강, 게임, 과학, 교육/학문, 금융/경제, 기관, 기업, 뉴스/미디어, 라이프스타일, 문화/예술, 사회, 성인, 쇼핑, 엔터테인먼트, 지역/여행, 컴퓨터/인터넷으로 분류하였다. 특히 이 중 컴퓨터/인터넷, 엔터테인먼트, 쇼핑, 게임, 교육 순으로 검색을 많이 한다고 발표하였다. 이 연구는 질의 유형 분류 시 질의뿐만 아니라 질의 이후 클릭로그를 검토하여 이용자의 검색 목적을 충분히 반영한 검색 질의 유형을 분류하였다는 것에 의의가 있다. 하지만 여전히 질의 분류를 위해 연구자가 직접 질의를 보고 주관적으로 판단한다는 점에서 한계가 있다.

3. 연구 방법

3.1 연구 개요

연구 프레임은 Fig. 1과 같다.

1) 데이터 수집

데이터 수집 단계에서는 검색엔진 DB로부터 검색 이용자의 검색 행동 로그를 수집하고 분석에 필요한 데이터마트를 구축한다. 데이터는 검색 질의별로 group by 되어 집계되며 검색 이후 검색 행동을 측정하는 정형 데이터와 검색 이후 클릭한 상품명인 비정형 텍스트데이터로 구성된다.

2) 데이터 전처리

데이터 전처리 단계에서는 토픽모델링에 활용할 비정형 텍스트데이터를 텍스트 마이닝, 자연어처리 과정을 통해 정제한다. 데이터 클렌징 작업 및 토큰화 작업 이후 단어-문서 단위인 Term-Frequency Matrix로 변환한다.

3) 검색 질의 주제 유형 분석

본 단계에서는 비정형 데이터에서 주제를 추출하는 토픽모델링 기법 중 하나로 가장 많이 사용되는 LDA(Latent Dirichlet Allocation) 알고리즘을 활용하여 검색 질의 주제를 정의한다. LDA 모델의 parameter tuning 작업을 통해 최적 parameter, 토픽 개수를 설정하고 실험을 진행한다. 검색 질의 주제 유형 정의 단계에서는 토픽모델링 결과로 얻어진 토픽 기반으로 유사한 단어로 구성된 문서(검색 질의)의 주제를 도출하여 검색 질의의 주제 유형을 분석하고 정의한다.

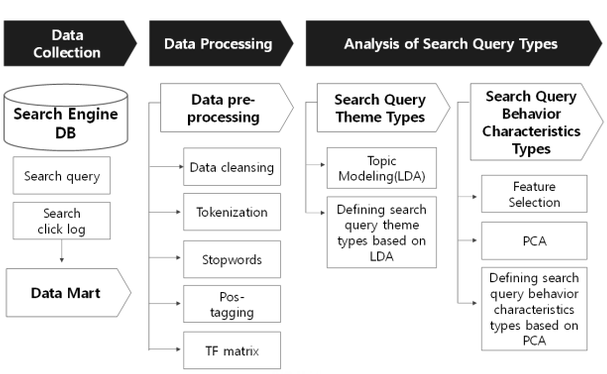


Fig. 1. Research Frame Work

4) 검색 질의별 검색 행동특성 분석

검색 질의 주제별 검색 행동을 대표하는 변수를 탐색하고 12개를 선정하여 주성분 분석을 진행하여 제 1, 2 주성분을 도출하고 주성분의 특성을 명명한다. 이후 두 주성분을 x축, y축으로 한 직교좌표평면에 15개의 검색 주제 유형을 투사하여 총 4개의 행동특성별 유형을 정의한다. 검색 질의 주제별 검색 행동특성 질의 유형을 분석하여 주제별로 유의한 검색 패턴의 차이가 존재하는지 파악한다.

3.2 이론적 배경

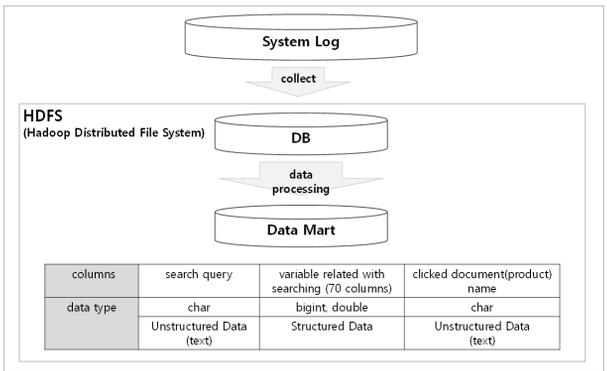
LDA(Latent Dirichlet Allocation)는 비지도학습을 수행하는 확률 모델(Generative probabilistic model)로서 주어진 문서에 대하여 Dirichlet 분포를 이용하여 단어별 토픽 분포와 문헌별 토픽 분포를 추론하는 토픽모델링 기법이다[8]. 토픽모델링 성능 측정 measure로는 perplexity(혼잡도), coherence score(응집성 지수)가 존재하는데 perplexity는 모델이 얼마나 토픽을 잘 나타내는지를 나타내기 때문에 토픽 개수가 증가할수록 perplexity 값이 개선된다. 그러나 해당 토픽들이 의미적으로 명확한 것, 즉 해석이 용이하다는 것을 의미하지는 않아 이를 보완하기 위한 지표로 coherence가 개발되었다. Coherence는 주제 내 단어의 유사도 계산 시 해당 단어와 주제의 핵심단어와의 의미적 유사도를 계산하여 의미론적으로 일치하는지를 파악하며 이를 통해 토픽의 일관성을 측정하게 된다. Coherence가 높을수록 토픽이 의미론적으로 일관성이 높다고 할 수 있다[9].

PCA(Principal Component Analysis)는 차원축소(Dimensionality Reduction Method) 기법으로 직교 변환을 이용하여 고차원 공간의 표본들을 선형 연관성이 없는 저차원 공간(주성분)의 표본으로 변환한다. 한 개의 축으로 사상시켰을 때 그 분산이 가장 커지는 축을 첫 번째 주성분(PC1)으로 설정하고 PC1과 직교하는 모든 방향 중 분산을 최대화하는 방향을 두 번째 주성분(PC2)으로 선정한다[10].

4. 데이터 수집 및 전처리

4.1 데이터 수집

데이터 원천은 국내 2019년 월평균 세션 수 2.6억 규모의 온라인커머스 사이트에서 2019년 1월~12월 1년간 발생한 검색 로그이다. 데이터 수집 프레임은 Fig. 2와 같다. 분석마트 구축 및 데이터 핸들링은 Hadoop에서 동작하는 Data Warehouse인 Apache Hive에서 HiveQL로 작업하였다. 데이터 스키마는 검색 질의별로 group by 하여 검색 행동특성 관련 변수와 토픽모델링을 위한 검색 이후 클릭 상품명 데이터를 수집하였다. 분석마트는 20만 행, 71열(200000x71)로 구성하였다.



columns	search query	variable related with searching (70 columns)	clicked document(product) name
data type	char	bigint double	char
	Unstructured Data (text)	Structured Data	Unstructured Data (text)

Fig 2. Data Acquisition Frame Work

분석 대상 검색 질의 선정 기준은 대표성 및 인기도에 초점을 맞추어 기간 내 누적 검색횟수 기준 상위 200,000개의 검색 질의를 수집하였다. 이는 연간 발생한 약 8,300만 개 검색 질의 중 0.24%를 차지하는 볼륨이며 검색횟수 비중으로는 61.6%를 차지하여 많은 검색을 유발하는 short-head 질의로 구성되어 있다. 예외처리한 검색 질의로 2019년도 말부터 유행한 질병인 코로나 관련 검색 질의가 있다. 해당 검색 질의가 특정 기간 내 급증하면서 예년과 구분되는 특수한 특징으로 토픽모델링 시 하나의 주제로 분류되어 범용적 검색 질의 주제 대비 대표성이 떨어지는 것을 방지하기 위해 '마스크', '세정제', '소독', '코로나'라는 단어가 포함된 검색 질의는 필터링하여 집계대상에서 제외하였다. 코로나 관련 검색 질의 특성에 대한 연구는 향후 연구 과제로 남기어 추가 분석하고자 한다. 또한 토픽모델링을 위해서 검색결과 클릭이 한 건이라도 발생한 검색 질의만이 클릭문서명이 남기 때문에 검색 이후 문서 클릭 결과가 0건인 질의는 필터링하여 집계에서 제외하였다. 검색 행동특성 관련 변수의 결측치에 대해서 0으로 대체하여 결측값을 처리하였다. 검색 이후 클릭 문서명(상품명) 관련하여서는 최대 50개의 문서를 공백 기준으로 연결하여 수집하였다.

4.2 데이터 전처리

본 단계에서는 토픽모델링에 활용될 텍스트데이터인 검색 이후 클릭 문서명에 대해 데이터 정제, 토큰화, 불용어 제거, 품사태깅 등의 과정을 거쳐 데이터를 전처리하였으며 최종적으로 토픽모델링의 입력값 단위인 Term-Frequency matrix를 생성하였다.

1) 데이터 정제

텍스트데이터인 문서명(클릭 상품명)에 대해 정규식을 사용하여 한글, 영어의 문자만 남기고 기타 특수문자는 모두 제거하여 데이터를 정제하였다. 정형 데이터인 검색 이후 특성 관련 변수 관련하여서는 수치 집계 후 결측치는 0으로 대체하였고 이상치에 대해서는 구조적으로 발생 가능한 값이며 해당 값 또한 검색 질의의 특성을 나타내 줄 수 있다고 판단하여 별도 처리를 하지 않았다.

2) 토큰화(Tokenization)

비정형 텍스트데이터인 문서명(클릭 상품명)에 대하여 공백 기준으로 데이터를 분리하는 토큰화 작업을 진행하였다. 해당 작업은 sklearn 라이브러리의 TfidfVectorizer 모듈을 활용하였다.

3) 불용어(Stopwords) 제거

텍스트데이터에서 유의미한 정보를 얻기 위해서 빈번하게 등장하지만 의미 분석을 하는데 거의 기여하지 않는 단어를 제거한다. 본 연구에서는 계량 관련 단어(kg, mm, ml 등), 조사(은, 는, 이, 가, 을, 를 등), 큰 의미 없이 빈번하게 쓰이는 단어(무료, 이벤트, 정품, 판매 등) 총 62개의 단어를 불용어로 간주하여 삭제하였다.

4) 품사태깅(Pos-tagging)

문서 내 단어들에 대해 품사를 태깅하여 유의한 의미를 내포하는 품사인 명사 키워드만 추출하였으며, POS-tagging은 python의 KoNLPy 패키지 Komoran 클래스를 사용하였다.

5) Term-Frequency Matrix

기계학습 시 비정형 데이터인 자연어를 컴퓨터가 연산할 수 있도록 백터(Vector)로 바꾸어 주는 작업 즉 워드 임베딩(Word Embedding)을 수행한다. TF-IDF(Term Frequency-Inverse Document Frequency)는 단어 빈도-역 문서 빈도로써 문서-키워드 매트릭스로 각 문서에 포함되어있는 키워드들의 빈도를 값으로 갖는 매트릭스이다. 해당 메트릭스는 토픽모델링의 입력값으로 활용된다.

5. 연구 결과

5.1 LDA를 통한 검색 질의 주제 유형 정의

1) 토픽 개수 최적화 실험

LDA 실험은 프로그래밍 언어로 Python을 활용하였고 IDE로 Jupyter Notebook(Anaconda3), 라이브러리로 gensim을 이용하였다. Table 1은 토픽모델링의 입력값 데이터 일부로 총 20만 개의 문서(검색 질의)별 검색결과 조회 문서명(클릭 상품명)을 기반으로 LDA를 수행하여 주제 도출 및 검색 질의의 주제 유형을 정의한다.

토픽모델링의 주요 파라미터는 토픽 개수 K를 선정하는 것인데 이를 위해 토픽 개수 최적화 실험을 수행하였다. 실험 평가 지표로는 토픽 간 의미론적 일관성을 더 잘 나타낼 수 있는 지표인 coherence(응집성 지수, c_v)를 택하여 토픽 개수를 늘려가며 지표의 score를 측정하였다. 실험 결과 Fig. 3과 같이 K=15에서 비교적 높은 coherence score 0.525가 측정되었고 40개 이상의 토픽 수는 유의미한 주제 그룹화에 용이하지 않다고 판단하여 토픽 수를 15개로 결정하였다.

Table 1. Dataset for Topic Modeling

Search Query	Clicked Document(product) Name
keyword1	나이키에어맥스 화이트 실버 검흰 올블랙 4종 나이키 스포츠웨어 테크 폴리스 (이하 생략)
keyword2	애플 에어팟 Apple Airpod MMEF2KH 제이스 엠식스 JAYS 무선이어폰 (이하 생략)
keyword3	베베숲 센시티브 80매 캡 10팩 비야비야 프로즌2 헤링본 물티슈 캡형 도리도리 (이하 생략)
keyword4	ASUS 정품파우치 랜젠더 뉴젠북 UX433FN-A6053T 인텔 i5-8265U 지포스 (이하 생략)
keyword5	커네스트 미포07플러스 블루투스이어폰 블랙에디션 듀얼DAC장착 (이하 생략)
keyword6	V2 세트 베이비백팁400g 투움바파스타 아웃백디지털상품권10만원권 (이하 생략)

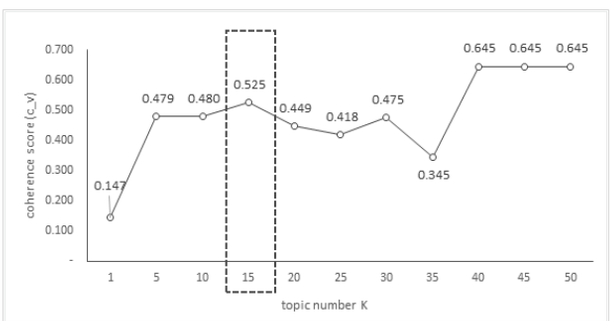


Fig. 3. Coherence Score According to the Number of Topics

2) 토픽 모델링 실험 및 토픽별 분포

토픽 수(num_topic) 파라미터를 15개로 최적화 후 Table 2와 같은 파라미터 설정으로 LDA 토픽모델링 실험을 진행하였다. 프로그래밍 언어는 Python을 사용하였고 gensim 라이브러리의 gensim.models.LdaMulticore 모듈을 사용하였다. 실험 결과로 얻은 coherence score는 0.49533이다.

Table 2. Parameter setting for Topic Modeling

Parameter	Setting Score
chunksize	2000
passes	20
iterations	1500
eval_every	None
coherence measure	C_v
num_topics	15

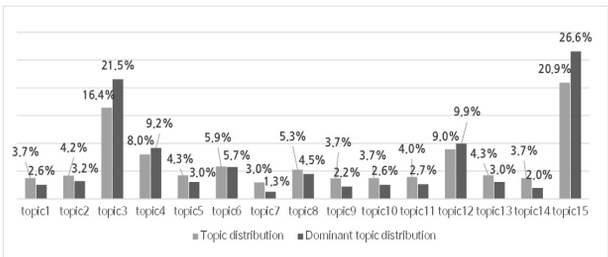


Fig. 4. Topic/Dominant Topic Distribution by Search Query

토픽모델링으로 도출된 15개의 토픽은 검색 질의 주제를 구성하는데 이에 대한 토픽별 분포를 Table 3과 같이 '검색 질의-Topic' matrix로 구성하여 분포를 살펴보았다. 또한 Table 4와 같이 검색 질의별로 가장 비중이 높은 topic에 1을 할당하고 그 외 topic은 0을 할당하여 가장 지배적인 topic의 분포는 어떠한지 살펴보았다.

검색 질의별 Topic/지배적 Topic의 분포를 Fig. 4와 같이 확인하였다.

Table 3. Search Query-topic Matrix

Topic / Search Query	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
나이키	0.01	0.01	0.76	0.01	0.01	0.01	0.01	0.01	0.13	0.01	0.01	0.01	0.01	0.01	0.01
에어팟	0.01	0.01	0.18	0.01	0.18	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.51
물티슈	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.89	0.01	0.01	0.01
노트북	0.01	0.67	0.01	0.01	0.01	0.25	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
블루투스이어폰	0.01	0.01	0.18	0.01	0.39	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.31
아웃백	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.39	0.02	0.34	0.02	0.02	0.02

Table 4. Search Query-dominant Topic Matrix

Topic / Search Query	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
나이키	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
에어팟	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
물티슈	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
노트북	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
블루투스이어폰	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
아웃백	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Topic_15가 topic 비중 20.9%, 지배적 topic 비중 26.6%로 가장 높았으며 다음으로 topic_3의 topic 비중 16.4%, 지배적 topic 비중 21.5%로 두 번째로 비중이 높았다. 후 순으로 지배적 topic 비중 기준 5%를 초과하는 topic은 topic_4, topic_6, topic_12이며 나머지 9개 topic은 비중이 5% 미만으로 확인되었다. Topic_15, topic_3과 같이 전체에서 높은 비중을 차지하는 topic의 경우 순수한 topic 분포보다 지배적 토픽으로 비중 계산 시에 대략 5%p의 큰 폭으로 비중이 더 높다. 이를 통해 검색 질의 내 특정 topic의 분포가 지배적으로 두드러진다기보다는 다양한 topic들로 구성되어 있으며 지배적 토픽 계산 시에 그중 연관도가 가장 높은 특정 topic에 집중되는 것을 파악하였다.

3) 토픽모델링 결과 기반 검색 질의 주제 정의

LDA 토픽모델링을 수행하여 15개의 토픽과 토픽별 가장 관련 깊은 키워드 결과 얻었고 이를 Table 5에 정리하였다.

Table 5. Definition of Search Query Topic Type Based on Topic Modeling Results

TOPIC	topic_1	topic_2	topic_3	topic_4	topic_5
Topic Definition / Topic Keyword	주방관련 식품/가전 (Kitchen related food/appliances)	컴퓨터/교육/유아동 (Computer/Education /Childhood)	패션의류 (Fashion)	가공/건강식품 (Processed/ Healthy Food)	생활가전/생필품 (Household appliances & necessities)
keyword1	생수	노트북	나이키	포	이어폰
keyword2	두유	학년	빅사이즈	개	다이슨
keyword3	코베아	그램	티셔츠	정	무선청소기
keyword4	전기레인지	수학	자켓	캡슐	겹
keyword5	구	ASUS	원피스	오뚜기	생리대
keyword6	팩	갤럭시워치	팬츠	팩	화장지
keyword7	베지밀	초등	아디다스	개월분	HDMI
keyword8	척테일러	레노버	여름	박스	수도꼭지
keyword9	그릴	국어	반팔	병	롤
keyword10	빌트인	너프	롱	정관장	크리넥스
TOPIC	topic_6	topic_7	topic_8	topic_9	topic_10
Topic Definition / Topic Keyword	라이프뷰티 (Life & Beauty)	취미 (Hobby)	라이프플러스 (LifePlus)	스마트디지털 (Smart Digital)	컴퓨터주변/e쿠폰 (Computer Peripheral/ e-Coupon)
keyword1	크림	레고	귀걸이	노트	GB
keyword2	로션	어벤져스	공기청정기	A	지포스
keyword3	헤라	아이언맨	목걸이	케이스	GTX
keyword4	에센스	마블	자동차	갤럭시	갤럭시탭A
keyword5	스킨	로봇청소기	반지	아이폰	갤럭시탭S
keyword6	클렌징	지프	키홀더	플러스	만원권
keyword7	설화수	피규어	팔찌	갤럭시S	RTX
keyword8	이니스프리	미밴드	순금	XS	D
keyword9	아이오페	레고호환	골프공	풀커버	인텔
keyword10	앰플	엔드게임	골드바	J	EOS
TOPIC	topic_11	topic_12	topic_13	topic_14	topic_15
Topic Definition / Topic Keyword	생필품/식품 (Necessities/Food)	트랜드잡화 (Trend miscellaneous goods)	레포츠/아웃도어 (Leports & Outdoors)	여행/성인 (Travel/Adult)	가구/인테리어 (Furniture&Interior)
keyword1	맥심	슬리퍼	시마노	강원	단
keyword2	하기스	백팩	다이와	단계	차량용
keyword3	기저귀	캐리어	등산화	리조트	LED
keyword4	카누	아쿠아슈즈	장갑	성인용품	원목
keyword5	제주	모자	스피닝릴	테일러메이드	인테리어
keyword6	코렐	크로스백	자전거	호텔	케이스
keyword7	단계	에코백	부산	진동기	다용도
keyword8	모카골드	슬링백	넥워머	경북	cm
keyword9	커피믹스	지갑	방한장갑	경주	A
keyword10	장우산	숄더백	모기퇴치기	앱솔루트	세대

15개의 토픽에 대하여 각각 토픽별 주요 연관 키워드들을 기반으로 토픽의 주제를 정의하였다. 검색 데이터 분석 이력이 있는 연구자의 판단에 근거하여 주제와 연관성이 높은 키워드들을 선별하여 최종적으로 토픽별 주요 키워드 10개를 선별하여 정리하였다.

Topic_1의 주요 연관 키워드는 '생수, 두유, 전자레인지, 그릴' 등으로 '주방 관련 식품/가전'으로 주제를 정의하였다. topic_2는 '노트북, ASUS, 레노버' 등 노트북 관련 키워드와 '학년, 수학, 초등', '국어, 너프' 등이 주요 연관키워드로 해당 topic은 '노트북/교육/유아동'으로 주제를 정의하였다. Topic_3의 주요 키워드는 '나이키, 빅사이즈, 티셔츠' 등 의류 관련 키워드가 주를 이루어 '패션의류'로 주제 정의하였다. Topic_4의 주요 키워드는 '포, 캡슐, 정관장' 등 건강식품 수량 관련 키워드와 '오뚜기' 등 가공식품 관련 키워드 등으로 '가공/건강식품'을 주제로 정의하였다. Topic_5의 주요 키워드는 '이어폰, 다이슨, 무선청소기' 등의 생활가전 키워드와 '생리대, 화장지' 등 생필품 관련된 키워드로 구성되어 있어 주제를 '생활가전/생필품'으로 정의하였다. Topic_6의 주요 키워드는 '크림, 로션', '헤라, 에센스' 등 뷰티용품 연관 키워드로 구성되어 '라이프뷰티'로 주제를 정의하였다. Topic_7의 주요 키워드는 '레고, 어벤져스, 아이언맨, 피규어, 앤드게임' 등의 취미 관련 키워드로 '취미'로 주제 정의하였다. Topic_8의 주요 키워드는 '귀걸이, 공기청정기, 반지, 팔찌' 등 생활에 플러스 되는 사치재 성격의 키워드로 구성되어 '라이프플러스'로 주제 정의하였다. Topic_9의 주요 키워드는 '노트, A, 케이스, 아이폰' 등 휴대폰 관련 키워드들로 구성되어 '스마트디지털'로 주제 정의하였다. Topic_10는 'GB, 지포스, GTX, 인텔' 등 컴퓨터 관련 키워드와 '만원권' 과 같은 상품권 관련 키워드로 구성되어 있어 '컴퓨터주변/e쿠폰'으로 주제 정의하였다. Topic_11의 주요 키워드는 '맥심, 하기스, 기저귀, 카누' 등으로 구성되어 '생필품/식품'으로 주제 정의하였다. Topic_12의 주요 키워드는 '슬리퍼, 백팩, 크로스백, 가방, 캐리어, 모자' 등으로 구성되어 있어 '트랜드잡화'로 주제 정의하였다. Topic_13은 '시마노, 다이와, 스피닝릴' 등 낚시 관련 키워드와 '등산화, 넥워머, 방한장갑' 등 아웃도어 관련 키워드들이 주요 키워드로 구성되어 '레포츠/아웃도어'로 주제 정의하였다. Topic_14는 '강원, 리조트, 호텔, 경부' 등 여행 관련 키워드와 '성인용품'과 같은 성인 관련 키워드가 주요 키워드로 구성되어 '여행/성인'으로 주제 정의하였다. Topic_15는 '단, 차량용, LED, 무선, 거치대, 접이식' 등의 키워드들이 주요 키워드로 구성되어 '가구/인테리어'로 주제 정의하였다.

5.2 PCA를 통한 검색 질의 주제별 검색특성 정의

토픽모델링으로 도출한 15개의 검색 주제 유형에 대해 검색 행동특성을 분석하여 검색 행동특성별로 유형을 구분하고자 한다. 이를 위하여 검색 행동특성 관련 주요 변수를 선정하고 토픽별로 변수 실적을 집계 후 변수 중 데이터의 특성을 가장 잘 구분 짓는 주요 변수를 추출하기 위하여 PCA(Principal Component Analysis)를 수행한다. PCA 결과 도출된 제1, 2 주성분 특성을 명명하고 최종적으로 두 주성분을 x축, y축으로 한 직교좌표평면에 15개의 검색 주제 유형을 투사하여 총 4개의 행동특성별 유형을 분석한다.

1) PCA 실험을 위한 변수 선택

PCA 실험을 위한 변수는 검색 로그 데이터 분석가의 정성적 평가로 Table 6과 같이 12개의 변수를 선정하였다. 12개의 변수는 검색서비스 이용 흐름에 따라 각 단계에서 주요하게 평가되어야 할 변수로 구성되어 있다. 검색 흐름 단계에 따라 12개 변수를 총 7개의 그룹으로 구분하였다. 7개의 그룹은 '검색 활성도, 검색결과 만족도, 검색 상품 만족도, 문서 클릭위치, 문서 클릭의 분산/집중도, 상품 가격, 광고상품 집중도로 구성하였다.

첫 번째 그룹은 '검색 활성도(Searching activity)'이며 이에 해당하는 변수는 '세션당 검색횟수'로 '검색횟수/세션수'로 계산된다. 두 번째 그룹은 '검색결과 만족도(Search

Table 6. Features Related to Onsite Searching Behavior

Category	Variables	Variable Logical Names
Searching activity	qc_per_sess	Query count per session
Search results satisfaction	ctr	Click-through rate(ClickCount/QueryCount)
Search product satisfaction	cc_per_prd	Click Count per product
	buy_try_ratio	Ratio of purchase attempts out of total Query Count
	buynow_ratio	Ratio of buy-now purchase attempts out of total purchase attempts
	qc_cvr	Order_Count/Query Count
Click position of document	clk_pos	Average document click position
	fst_clk_pos	First Clicked document position
Variance/Concentration of document click	m_clk_ratio	Ratio of most-clicked document Click Count out of total Click Count
	avg_prd_clk_ratio	Average click share by document
Price of search product	amt_per_ord	Average payment amount per order
Concentration of ad document	ad_qc_ratio	Ratio of searches that include advertisements among all searches

Table 7. Importance of PCA Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	2.736	1.442	0.947	0.782	0.708	0.457	0.362	0.217	0.140	0.111	0.081	0.030
Proportion of Variance	0.624	0.173	0.075	0.051	0.042	0.017	0.011	0.004	0.002	0.001	0.001	0.000
Cumulative Proportion	0.624	0.797	0.872	0.923	0.964	0.982	0.993	0.997	0.998	0.999	1.000	1.000

results satisfaction)' 이고 해당 변수는 검색 이후 검색결과 내에서 문서(상품)을 얼마나 클릭했는지 관련된 '클릭률(CTR, 문서클릭수/검색횟수)'로 선정하였다. 세 번째 그룹은 '검색 상품 만족도(Search product satisfaction)'이며 해당 변수는 검색결과에서 상품을 클릭 후 상품 만족도를 나타낼 수 있는 '상품당 클릭수(문서클릭수/unique 상품클릭수)', '검색 후 구매시도율(주문시도수/검색횟수)', '구매시도 중 바로구매 비중(바로구매 클릭수/구매시도횟수)', '검색 후 구매 전환율(검색 기여 주문수/검색횟수)'로 구성하였다. 네 번째 그룹인 '문서(상품) 클릭위치(Click position of document)'는 '평균 문서 클릭위치', '첫 문서 클릭위치' 변수로 구성하였다. 해당 지표에서는 검색 질의별로 이용자가 비교적 검색결과 상단 위치한 상품을 클릭하는지, 아니면 스크롤을 내려 더 하단의 상품을 탐색하고 클릭하는지 등의 검색 행동특성을 파악할 수 있다. 다섯 번째 그룹은 '문서(상품) 클릭 분산/집중도(Variance/Concentration of document click)'로 해당하는 변수로는 '최다클릭 문서 클릭점유율', '문서별 평균 클릭 점유율'이 있다. 여섯 번째 그룹은 '상품가격(Price of search product)' 이며 해당 변수로는 '주문당 평균 결제금액' 이다. 이는 검색을 통하여 결제까지 이루어진 경우 해당 주문의 평균적인 결제금액을 나타내는 지표이다. 마지막 일곱 번째 그룹은 '광고상품 집중도(Concentration of ad document)'로 해당 변수는 '검색광고 비중(문서 포함 검색/검색횟수)'이다. 검색결과에 리스팅되는 문서(상품)은 크게 광고 구좌에 리스팅 되는 광고상품과 검색랭킹 로직에 의해 리스팅되는 상품으로 나뉜다. 광고상품이 리스팅 되는 검색 질의는 주로 대중들의 검색량이 많은 인기 있는 질의인 경우이다. 해당 정보를 파악하기 위해 검색횟수 중 광고상품이 포함된 검색의 비중을 변수로 구성하였다.

2) PCA 수행

PCA 수행은 통계분석 소프트웨어인 R의 prcomp, FactoMineR 패키지를 사용하였다. PCA 수행 시 변수 간 선형결합을 유도할 때 분산을 이용하기 때문에 변수 scale에 영향을 받을 수 있어 이를 제거하기 위하여 계산 시 공분산행렬(covariance matrix) 대신 상관계수행렬(correlation matrix)를 사용하였다. PCA 수행결과는 하기 Table 7, Fig. 5와 같다.

PCA 수행결과를 기반으로 몇 개까지의 주성분을 사용할 것인지 결정한다. 주성분 선택을 위해 PCA의 분산(Standard deviation)인 고윳값(eigenvalue)이 1 이상인지를 확인하고

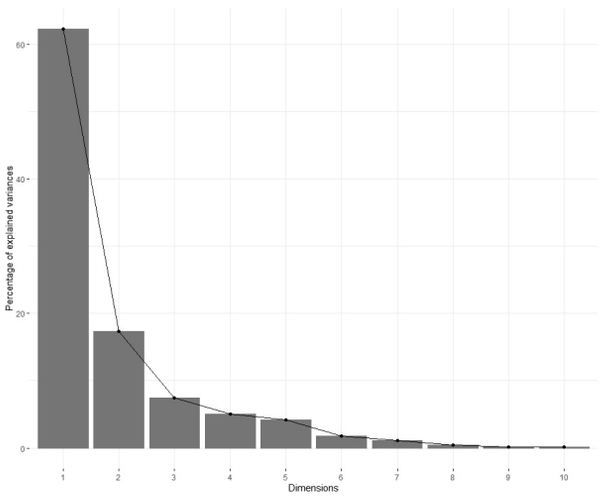


Fig. 5. PCA Scree Plot

총 변수의 누적 설명력이 80% 이상이 되는 성분까지를 선택 범위로 고려한다. Table 7을 참고하면 고윳값(Table 7 Standard deviation 표기)이 1이 상인 주성분(Principal Component)은 PC1, PC2이며 PC3은 0.947로 1에 근사한 수치이다. 두 번째 기준으로 주성분의 누적 비중(Table 7 Cumulative Proportion 표기)을 살펴보면 PC1만으로 전체 데이터 분산의 62.4%를 설명할 수 있고 PC2를 추가하면 79.7%를 설명할 수 있다. Fig. 5를 참고하면 주성분별 설명 가능 분산의 비중이 PC2 이후 급격히 완만해지는 모습을 확인할 수 있다. PC1, PC2 각각 고윳값이 1이 넘고 PC2까지의 누적 분산량이 주성분 채택 기준인 80%에 근사하여 PC1, PC2 기준 검색 행동특성을 살펴보기로 한다.

3) PCA 결과 분석

PC1, PC2가 어떠한 특성을 갖는지 파악하기 위해 각각 PC1, PC2에 영향을 미치는 변수가 무엇인지 살펴보았다. 하기 Table 8은 PCA 결과 PC6까지의 주성분별 변수 기여도이다.

Table 8의 기울임 처리된 부분은 PC와 주요하게 양의 관계인 변수이며 밑줄 처리된 부분은 PC와 주요하게 음의 관계에 있는 변수이다. PC1은 세션당 검색횟수가 높을수록, 클릭 위치가 낮을수록 양의 상관관계가 강하게 형성되며 음의 상관관계는 바로 구매하기 시도를 적게 할수록, 문서 집중도가 낮을수록 강하게 형성된다. 결론적으로 PC1은 검색 및 문서 탐색의 활성도를 나타내는 성분으로 특정 지을 수 있다. PC2는 검색결과에서 동일 문서를 여러 번 클릭할수록, 상품 가격

Table 8. Contribution of PCA Variables by Principal Component

Variables	PC1	PC2	PC3	PC4	PC5	PC6
qc_per_sess	***0.284***	- 0.040	0.441	- 0.241	0.522	- 0.070
ctr	0.255	0.204	- 0.201	- 0.633	- 0.485	0.291
cc_per_prd	- 0.165	***0.520***	0.324	- 0.132	- 0.246	- 0.681
buy_try_ratio	**- 0.312**	- 0.217	0.154	- 0.424	0.150	0.154
buynow_ratio	**- 0.347**	- 0.095	0.113	- 0.254	0.001	- 0.143
qc_cvr	**- 0.312**	- 0.282	0.114	- 0.362	- 0.064	0.032
clk_pos	***0.354***	0.019	0.145	- 0.049	- 0.213	- 0.056
fst_clk_pos	***0.351***	- 0.063	0.158	0.093	- 0.227	0.054
m_clk_ratio	**- 0.346**	0.118	- 0.175	0.083	- 0.161	0.116
avg_prd_clk_ratio	- 0.313	0.286	- 0.251	0.147	0.056	0.082
amt_per_ord	- 0.145	***0.483***	0.542	0.136	0.019	0.613
ad_qc_ratio	- 0.154	**- 0.466**	0.420	0.303	- 0.528	0.017

Table 9. Attribute Definition of PC1, PC2

Attribute Definition		Details
PC1	Search/Document browsing activity (high-searching /low-searching)	There is a deeply connected with click position in positive way and Variance/Concentration of document click in negative way.
		The more you click at the bottom of your search results(the more you browse), the less focused your documents and the percentage you buy-now.
PC2	Product Involvement (high-involvement /low-involvement)	It is deeply connected with cc_per_prd and amt_per_ord in positive way.
		When you click on the same document(product) multiple times in the search results, the unit price tends to be high.

높을수록 상관관계가 높아지는 경향이 있어 상품 관여도를 나타내는 주성분으로 특정 지을 수 있다. 위 분석을 바탕으로 PC1, PC2 특성을 다음 Table 9와 같이 '검색/문서 탐색 활성도', '상품 관여도'로 정의하였다.

정리하면 PCA를 통해 데이터의 약 80%를 설명할 수 있는 주성분 2개를 추출하였고 각각 제1 주성분, 제2 주성분에 큰 비중으로 결합된 변수들의 특성을 분석하여 주성분의 특성을 정의하였다. 제1 주성분은 총 분산의 약 62.4%를 설명하며 '검색/문서 탐색 활성도'를 나타낸다. 제2 주성분은 총 분산의 약 17.3%를 설명하며 '상품 관여도'를 나타낸다.

4) 질의 토픽별 검색 행동특성별 유형 정의

PCA를 통해 얻은 두 주성분을 바탕으로 Fig. 6과 같이 토픽의 검색 행동특성을 4가지로 분류했다. 2차원 평면에 x축은 PC1(검색/문서 탐색 활성도), y축은 PC2(상품 관여도)로 설정하여 각 성분이 0이 되는 지점을 기준으로 4개의 영역을 구분하였다.

Fig. 6의 제1 사분면은 검색/문서 탐색 활성도가 높고 상품관여도 또한 높은 '고관여 탐색형' 유형으로 정의한다. 해당 유형은 다수의 검색과 다양한 문서(상품) 클릭을 발생시키며 검색 이후 구매하는 상품의 가격이 고단가 또는 상품당 클릭 수가 높은 특성을 보인다. 제2 사분면은 검색/문서 탐색 활성도는 비교적 낮지만 상품관여도는 높은 '고관여 목적형'

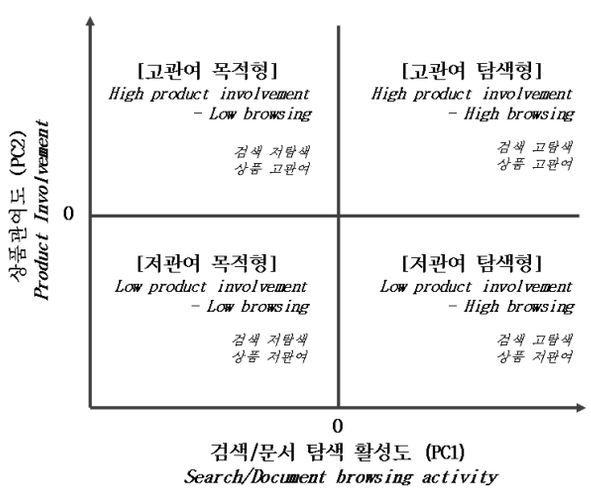
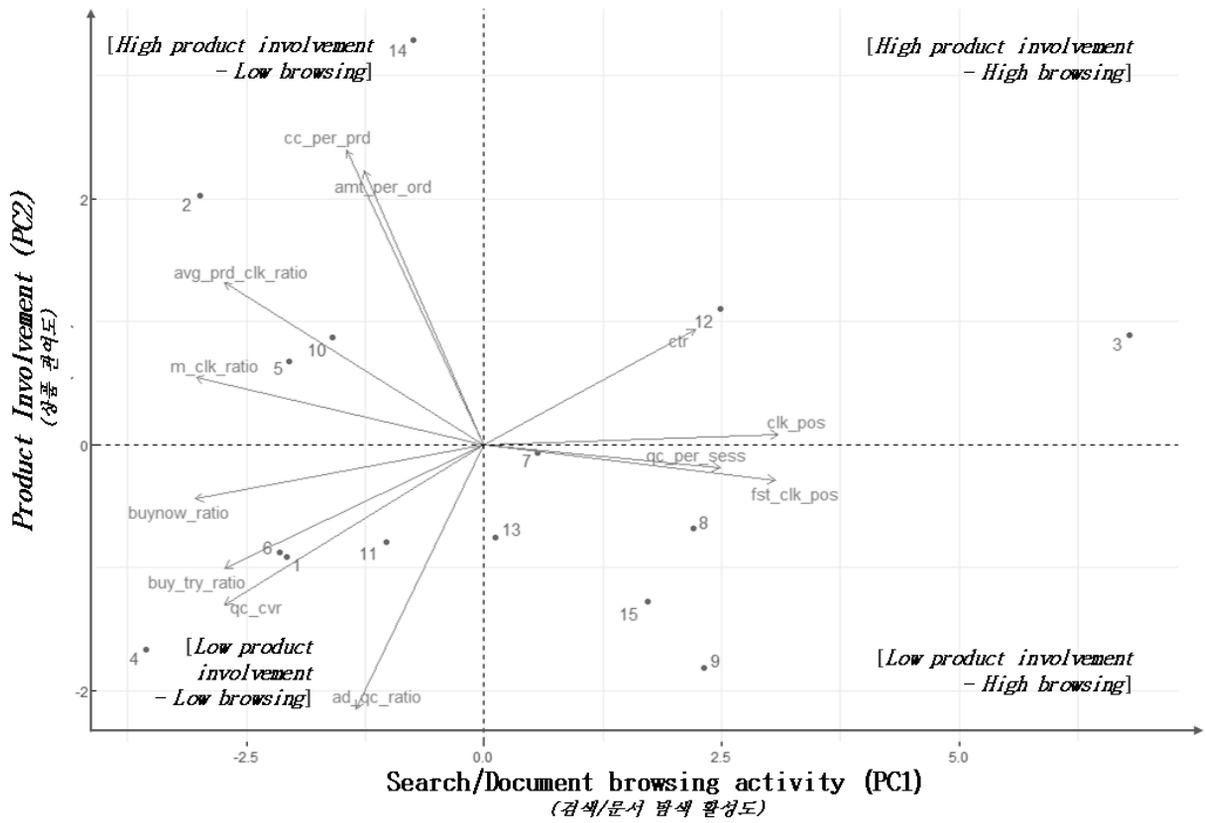


Fig. 6. Search Behavior Properties Type Definition

으로 정의한다. 해당 유형은 고가 상품을 구매하면서도 낮은 검색 활성도를 보여 탐색하는 상품에 대한 기준이 뚜렷한, 즉 목적성이 뚜렷하다고 판단하였다. 제3 사분면은 검색/문서 탐색 활성도, 상품관여도가 모두 낮은 '저관여 목적형'으로 소극적인 검색 활동을 보이며 특정 상품에 대한 관심도도 낮은 행동을 보인다. 해당 유형은 딱히 목적 없이 브라우징 목적으로 검색 활동을 하다 진입장벽이 낮은 저가의 상품을 충동적으로 구매하는 유형으로 판단된다. 제4 사분면은 '저관



Type of search behavior	Search Query Topic
High product involvemeat – High browsing (고관여 탐색형)	[3]패션의류(Fashion), [12]트랜드잡화(Trend miscellaneous oods)
High product involvemeat – Low browsing (고관여 목적형)	[2]컴퓨터/교육/유아동(Computer/Education/Childhood), [5]생활가전/생필품(Household appliances&necessities), [10]컴퓨터주변/e쿠폰(Computer Peripheral/e-Coupon), [14]여행/성인(Travel/Adult)
Low product involvemeat – Low browsing (저관여 탐색형)	[7]취미(Hobby), [8]라이프플러스(LifePlus), [9]스마트디지털(Smart Digital), [13]레포츠/아웃도어 (Leports&Outdoors), [15]가구/인테리어(Furniture&Interior)
Low product involvemeat – High browsing (저관여 목적형)	[1]주방관련 식품/가전(Kitchen related food/Home appliances), [4]가공/건강식품(Processed/Healthy Food), [6]라이프뷰티(Life&Beauty), [11]생필품/식품(Necessities/Food)

Fig. 7. Search Query Topic Projection on PC1, PC2 2D Biplot

여 탐색형'으로 검색/문서 탐색 활성도는 높지만 상품 관여도는 낮은 유형으로 정의한다. 해당 유형은 활발한 검색, 상품 탐색을 하며 특정 지배적인 상품을 클릭하는 것이 아니라 다양의 상품을 클릭하는 경향이 있다.

검색 행동특성 유형을 4가지로 정의 후 Fig. 7과 같이 각 질의 주제 유형을 두 주성분을 x축, y축으로 나눈 직교좌표 평면에 투사하여 어느 검색 행동특성 유형에 속하는지 분석하였다. Fig. 7을 참고하면 x축 PC1(검색/문서 탐색 활성도), y축 PC2(상품관여도)인 2차원 평면 위 토픽모델링을 통해 도출한 15개의 토픽 값을 투사하였다.

검색 행동특성 분류 유형 중 '고관여 탐색형'에 해당하는 토픽은 [3]패션의류, [12]트랜드잡화로 나타났다. 패션 관련된 검색 질의는 사용자가 찾고자 하는 상품에 대한 속성을 키워드에 적절히 반영하여 정보검색(information retrieval)이 이루어지는 것이 어렵기 때문에 보통 여러 번의 검색과 문서 클릭 등의 탐색 패턴을 보이는 것으로 사료된다. 이와 같은 특성이 반영되어 해당 유형의 토픽들은 세션당 검색횟수가 많고 탐색활동을 활발히 하여 검색결과 리스팅 하단의 상품을 클릭하는 경향이 있다. 탐색이 활발한 만큼 반대로 검색횟수당 구매시도율이나 구매전환율 등은 낮은 편이다. 또한 클릭한 개별 상품에 대한 집중도가 높은 편이다.

'고관여 목적형'에 해당하는 토픽은 [2]컴퓨터/교육/유아동, [5]생활가전/생필품, [10]컴퓨터주변/e쿠폰, [14]여행/성인으로 나타났다. '고관여 목적형' 유형은 '고관여 탐색형' 대비 검색 탐색 활성도 관련하여 더 적은 검색과 비교적 특정 문서(상품)에 집중된 클릭, 검색결과 상단의 상품을 클릭하는 등의 소극적 탐색활동 특징을 보인다. 이 유형은 검색결과 상단에 위치한 특정 인기상품 또는 검색 랭킹이 높은 상품이 검색 전체 문서 클릭에서 차지하는 비중이 높은 경우가 많다. 이는 토픽의 주제와 비교하여 해석하면 더욱 명확한데, '컴퓨터', '가전', '쿠폰', '여행' 주제 관련 키워드는 사용자가 질의에 정보요구를 비교적 구체적으로 담을 수 있어 사용자가 검

색엔진에 질의 제시 후 의도하던 정보를 잘 포함하고 있는 적합문서(relevant document)를 발견할 가능성이 크다. 즉 정보검색 시 주제 적합성(topical relevance)과 사용자 적합성(user relevance)을 모두 충족시킬 확률이 높다.

‘저관여 탐색형’에 해당하는 토픽은 [7]취미, [8]라이프플러스, [9]스마트디지털, [13]레포츠/아웃도어, [15]가구/인테리어로 나타났다. 해당 유형에 속하는 토픽은 검색/문서 탐색 활성도는 높으나 상품 관여도가 낮으며, 특정 상품이 인기 있는 것이 아니라 여러 상품이 클릭을 받는 패턴을 보인다.

‘저관여 목적형’에 해당하는 토픽은 [1]주방관련 식품/가전, [4]가공/건강식품, [6]라이프뷰티, [11]생필품/식품으로 나타났다. 해당 유형에 속하는 토픽은 검색/문서 탐색 활성도도 낮으며 저관여 상품에 해당한다. 다양한 탐색활동 등을 하지 않고 특정 지배적인 문서(상품)이 존재하는 경우가 많으며 검색결과 내 특정 상품에 대한 중복 클릭이 상대적으로 적은 특성을 보인다. 해당 유형은 딱히 목적 없이 브라우징 목적으로 검색 활동을 하다 진입장벽이 낮은 인기상품을 충동적으로 구매하거나 과거 구매했던 상품을 적은 탐색 과정을 거쳐 반복구매하는 유형으로 판단된다.

6. 결 론

본 논문은 검색 질의 연구에 빅데이터를 활용한 기계학습 기반 연구 방법론을 제안하였으며 이를 기반으로 검색 질의 주제 분류 관련 의미 있는 결과를 도출하였다. 온라인커머스 쇼핑몰에서 1년간 발생한 검색로그를 바탕으로 20만 개의 검색 질의에 대한 정형, 비정형 데이터를 수집하였고 비정형 텍스트데이터 기반 토픽모델링 LDA를 수행하여 coherence score 0.49533 수준으로 15개의 토픽과 토픽별 주요 키워드를 도출하였다. 도출된 15개의 토픽에 대해 주요하게 영향을 미치는 영향도 상위 키워드를 고려하여 15개의 검색 질의 주제 유형을 정의하였다. 검색 질의 주제 유형 15개는 ‘주방관련 식품/가전’, ‘노트북/교육/유아동’, ‘패션의류’, ‘가공/건강식품’, ‘생활가전/생필품’, ‘라이프뷰티’, ‘취미’, ‘라이프플러스’, ‘스마트디지털’, ‘컴퓨터주변/e쿠폰’, ‘생필품/식품’, ‘트랜드잡화’, ‘레포츠/아웃도어’, ‘여행/성인’, ‘가구/인테리어’로 정의하였다. 본 연구는 선행 연구에서 정성적 방법과 제한적인 데이터로 진행되었던 검색 질의 분류 연구에 기계학습 방법을 적용하고 빅데이터를 활용하여 데이터 기반 객관성을 확보한 연구 방법론은 제안하였다는데 의의가 있다.

또한 검색 질의 유형 관련 새로운 분류체계를 제시하였다. 검색 질의 자체가 갖는 의미뿐만 아니라 검색 행동특성을 반영하여 검색 질의를 분석할 수 있도록 검색 질의별 12개의 검색 행동특성 변수를 선정하고 PCA를 수행하여 2개의 주성분을 도출하였다. 제 1, 2 주성분 특성을 파악하고 최종적으로 두 주성분을 x축, y축으로 한 직교좌표평면 기준으로 4개의 검색 행동특성 유형 ‘고관여 목적형’, ‘고관여 탐색형’, ‘저관여 목적형’, ‘저관여 탐색형’을 정의하였다. 이후 15개의 검색 주제 유형을 좌표 평면에 투사하여 검색 질의 주제별로 어떠한 검색 행동특성을 가지는 복합적으로 분석하였다. 이를 통해 검색 로그 대상 통계적인 분석 방법론을 적용하여 기존 연구에서 다뤄지지 않았던 이용자의 검색 패턴 특성을 고려한 검색 질의 분류 체계를 제안하였다는 점에서 해당 연구 분야의 연구 방법론과 주제의 다양성을 제고하였다. 본 연구는 효과적인 검색서비스 구축 및 검색 시스템 개발에 기여할 것으로 기대된다.

본 연구를 발전시킨다면 기계학습 텍스트분류기를 모델링하여 주어진 검색 질의 관련 데이터 기반으로 검색 주제 유형을 자동으로 분류하는 분류기를 개발할 수 있을 것으로 기대되며 이는 추후 연구과제로 남긴다.

References

[1] H. I. Kwon, B. H. Baek, Y. J. Ahn, and J. H. Lee, “A Study on the Development Strategies for e-commerce Innovation,” *Journal of the Korea Contents Association*, Vol.20, No.1, pp.217-232, 2020.

[2] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, “Analysis of a very large web search engine query log,” *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, Vol.33, No.1, pp.6-12, 1999.

[3] A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic, “Searching the web: The public and their queries,” *Journal of the American Society for Information Science and Technology*, Vol.52, No.3, pp.226-234, 2001.

[4] A. Spink, B. J. Jansen, D. Wolfram, and T. Saracevic, “From e-sex to e-commerce: Web search changes,” *IEEE Computer*, Vol.35, No.3, pp.133-135, 2002.

[5] B. J. Jansen, A. Spink, and J. Pedersen, “A temporal comparison of Alta Vista web searching,” *Journal of the American Society for Information Science and Technology*, Vol.56, No.6, pp.559-570, 2005.

[6] NCM. Ross and D. Wolfram, “End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine,” *Journal of the American Society for Information Science and Technology*, Vol.51, No.10, pp.949-958, 2000.

[7] S. Y. Park, J. H. Lee, and J. S. Kim, “Analysis of Query Types and Topics Submitted to Naver,” *Journal of the Korea Society for Library and Information Science*, Vol.39, No.1, pp.265-278, 2005.

[8] S. Y. Bong and K. B. Hwang, “Applying Labeled LDA to Author Keywirds Recommendation,” in *Proceedings of KIISE Spring Conference*, pp.385-389, 2010.

[9] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, “Automatic evaluation of topic coherence,” in *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp.100-108, 2010.

[10] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, Vol.24, No.6, pp.417-441, 1933.

강 현 아

https://orcid.org/0000-0002-6666-5000
e-mail : khariss@korea.ac.kr
2015년 성신여자대학교
미디어커뮤니케이션/경영학과
(커뮤니케이션/경영학사)
2019년 ~ 현 재 고려대학교
빅데이터융합학과 석사과정
관심분야 : 자연어처리, 기계학습, 인공지능, 텍스트마이닝

임 희 석

https://orcid.org/0000-0002-9269-1157
e-mail : limhseok@korea.ac.kr
1992년 고려대학교 컴퓨터학과(이학학사)
1994년 고려대학교 컴퓨터학과(이학석사)
1997년 고려대학교 컴퓨터학과(이학박사)
2008년 ~ 현 재 고려대학교 컴퓨터학과
교수
관심분야 : 자연어처리, 인공지능, 기계학습, 뇌신경언어정보처리