유사한 주제의 카테고리를 가진 국내 온라인 커뮤니티 간 성향 분석

2조 윤여수, 함다희, 코작바로바무니사

1. 연구 목적 및 절차

연구 목적

유사한 주제의 카테고리를 가진 여러 커뮤니티 간의 성향에 차이가 있는지를 분석해보고, 사용자가 원하는 정보에 맞는 커뮤니티에 수월하게 접근하여 더욱 좋은 질의 정보를 얻게 한다.

절차

데이터 마이닝의 과정에 따라 연구를 진행한다.

2. 연구 진행

카테고리 및 웹사이트 선정

카테고리는 임의로 "운동", "패션", "축구" 세 주제로 정한다. 웹사이트는 전 세계의 웹사이트를 국가별로 순위를 매겨주는SimilarWeb에서 상위 30위권에 속해있는 웹사이트를 선정한다.

세 사이트는 각각 7위, 20위, 26위를 차지하고 있다.

SimilarWeb

Rank	Website
7	dcinside.com
20	fmkorea.com
26	ppomppu.co.kr

그림 1. 웹사이트 순위

웹크롤링

카테고리의 게시글을 20,000개씩 크롤링 하였고 패션 카테고리만 데이터의 양이 부족하여 10,000개씩 크롤링 하여 데이터를 수집하였다.

장서빈도(Collection Frequency) 분석

빈도수를 구하여 그림 2와 같이 CF값을 계산한다.

그림 2. CF 계산 과정

수집된 데이터는 각 커뮤니티의 카테고리별로 글 번호, 제목, 본문, 날짜로 나누어 CSV파일로 저장한다.

웹사이트의 특성상 인터넷 텍스트가 많고, 웹크롤링 과정에서 모은 데이터의 양이 많으므로 KoNLPy에서 제공하는 Okt를 사용한다.

불용어 처리

CF를 분석한 CSV파일에서 상위 50개 단어를 위주로 자모음과 숫자를 제거한다. 다음으로는 1 음절 단어를 찾아 의미 없는 단어를 제거하고, 2 음절 이상의 단어에서 복합명사나 의미가 없는 단어들을 을 제거한다.

또한 단순빈도 분석에서 검색어는 의미가 없으므로 카테고리의 제목을 추가적으로 제외하였다.

코사인 유사도 분석

불용어를 제거한 상위 50개 단어를 식(1)을 이용하여 유사도 분석을 한다.

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

식 (1)

코사인 유사도 분석은 많은 차원 공간에서의 거리 측정이 가능하므로 본 연구에서 사용한 게시글 데이터를 분석하는데 알맞았다.

3. 분석 결과

CF 데이터 분석 결과

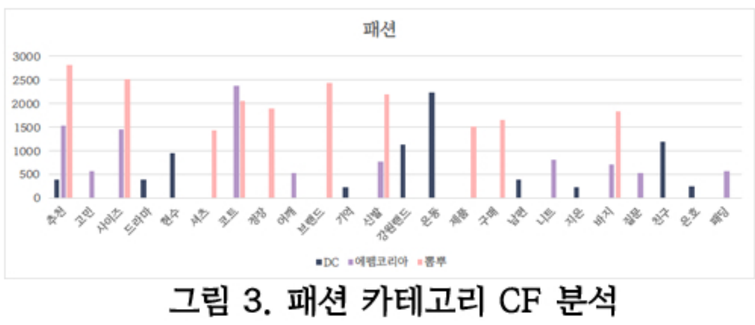

그림 3. 패션 카테고리 CF 분석

DC의 경우 당시 방영되었던 드라마에 대한 이야기가 많았고, 다른 두 커뮤니티는 전체적으로 아이템과 패션 추천에 관련된 단어들이 주를 이루었다.

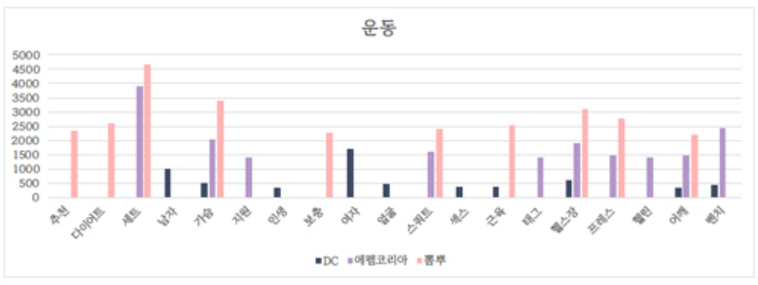

그림 4. 운동 카테고리 CF 분석

DC의 경우 성(性)과 관련된 내용이 반을 차지 하였고 다른 두 커뮤니티는 운동 부위나 운동 종류와 관련된 내용이 많이 나왔다.

그림 5. 축구 카테고리 CF 분석

에펨코리아는 수능과 관련된 데이터가 상위권에 위치해 있으며 DC는 손흥민, 메시 등 인물과 관련해서 많이 나왔고, 뽐뿌는 인물, 경기, 감독 등 편향되지 않고 골고루 분석이 되었다.

코사인 유사도 분석 결과

(a)

Fashion	뽐뿌	DC	에펨코리아
뽐뿌	0	0.05	0.57
DC	0.05	0	0.04
에펨코리아	0.57	0.04	0

(b)

Health	뽐뿌	DC	에펨코리아
뽐뿌	0	0.28	0.70
DC	0.28	0	0.29
에펨코리아	0.70	0.29	0

(c)

Soccer	뽐뿌	DC	에펨코리아
뽐뿌	0	0.43	0.41
DC	0.43	0	0.29
에펨코리아	0.41	0.29	0

그림 6. 코사인 유사도 분석

축구 카테고리를 제외한 다른 두 카테고리에서는 디시인사이드를 제외한 두 커뮤니티 사이트의 유사도가 상대적으로 높게 나왔으며 DC와 에펨코리아의 유사도는 전체적으로 높지 않음을 보아 두 커뮤니티간 성향이 같지 않다는 것을 확인할 수 있다.

4. 결 론

패션 카테고리

에펨코리아와 뽐뿌는 게시글의 유사도가 상대적으로 비슷하며 패션 추천에 관련된 내용이 주를 이루었다. 다만 디시인사이드의 경우 드라마에 관련된 내용이 주를 이루었음을 확인하였다.

운동 카테고리

에펨코리아와 뽐뿌는 게시글의 유사도가 높으며 운동 부위, 종류, 추천 등의 글들이 많았고, 디시인사이드는 성(性)과 관련된 내용이 주로 차지함을 확인하였다

축구 카테고리

뽐뿌는 인물, 팀, 리그 등 골고루 분석이 되었고, DC는 인물 위주로 분석이 되었으며 에펨코리아의 상위 데이터는 수능위주와 관련된 단어로 이루어져 있었다. 대체적으로 전체적인 유사도가 낮으며 특히 DC와 에펨코리아와의 유사도가 더 낮은 것을 보아 성향이 다름을 확인하였다.