소셜 네트워크 재검색 시스템의 설계

심규리[O], 김동현[*]
[O]동서대학교 소프트웨어학과,
[*]동서대학교 소프트웨어학과
e-mail: andand4306@gmail.com, pusrover@dongseo.ac.kr

The Design of Rescreening System for Social Network

Gyu Ri Sim[O], Dong Hyun Kim[*]
[O]Department of Software, Dongseo University,
[*]Department of Software, Dongseo University

● 요 약 ●

최근 소셜 네트워크 서비스 시장이 급속히 성장함에 따라 SNS 사용자 또한 지속적으로 증가하고 있다. 그러나, 광고성 게시물도 함께 증가함에 따라 해시태그 기반 검색의 정확도가 감소하는 문제점을 가지고 있다. 본 연구에서는 SNS 검색 활동의 정확도와 효율성을 개선하기 위하여 SNS 해시태그 기반 재검색 시스템을 제안한다. 제안 시스템을 적용하면 SNS 사용자의 검색 활동의 정확도와 효율성이 증가할 것으로 기대된다.

키워드: 키워드(KeyWord), 코사인유사도(Cosine Similarity), 코엔엘파이(KoNLPY) 해시태그(HashTag)

I. 서론

최근 소셜 네트워크 서비스(SNS, Social Network Service)로의 접근성이 향상됨에 따라, 사용자들은 SNS를 통해 관심 분야 및 일상생활의 내용을 공유하고 있다. SNS 사용자들이 일상생활을 공유함에 따라, 사용자들이 알고 싶은 주제를 나타내는 편의성을 높인 '#맛집', '#카페' 등과 같이 #(해시 기호) 뒤에 단어를 붙여 검색하는 해시태그(Hashtag) 검색 활동도 증가하게 되었다.

현재의 해시태그 검색 활동은 해시태그 검색어와의 일치성, 조회 수 등을 이용하여 검색 결과의 순서를 결정하고 있다. 해시태그 검색을 통해 정보를 얻고자 하는 사용자들에게는 광고성의 게시물이 상단에 위치하게 되는 검색의 정확도와 효율성이 감소하는 문제점들이 있다. 이 점을 악용하여 게시물을 작성하는 사용자는 해시태그 검색어와 작성한 글의 관련성이 전혀 없는 일명 광고성의 게시물에 해시태그를 무분별하게 사용하고 있다.

본 연구에서의 관련 연구로는 기존의 키워드 추출[2] 및 유사도 분석[3] 기능이 있다.

이에 본 연구에서는 위의 문제점들을 개선하기 위하여 결과 내에서 해시태그 검색어와 맞지 않는 게시물들의 내용을 분석 후 재검색을 통하여 검색 결과를 재정렬하는 SNS 해시태그 재검색 시스템을 제안한다.

논문의 구성은 다음과 같다. 2장에서는 관련 연구를 기술하고 3장에서 재검색 시스템의 설계를 제안한다. 마지막으로 4장에서 결론을 기술한다.

II. 관련 연구

본 연구에서 관련 연구로 첫 번째는 키워드 추출[2]이 있다. 키워드 추출은 문서의 내용을 대표할 수 있는 단어를 추출하는 과정이다. 키워드 추출은 문서에서 키워드 후보군을 생성하고, 생성된 키워드 후보군의 단어들에 따라 점수를 매기고 점수가 높은 후보군을 키워드로 추출한다. 두 번째로는 유사도 분석[3]이 있다. 유사도 분석은 두 벡터(문자열) 간의 유사도를 측정하는 사용되는 것이다. 두 벡터(문자열) 간의 각도를 이용하여 구할 수 있는 유사도를 의미한다. 두 벡터의 방향이 완전히 동일한 경우에는 1의 값을 가지며, 90°의 각을 이루면 0, 180°로 반대의 방향을 가지면 -1의 값을 가지게 된다. 즉, -1 이상 1 이하의 값을 가지며 값이 1에 가까울수록 유사도가 높다고 판단할 수 있다. 이는 두 벡터가 가리키는 방향이 얼마나 유사한가를 의미한다.

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Fig. 1. 코사인 유사도 식

III. 해시태그 검색어 선별 시스템 설계

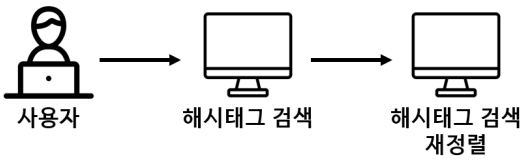

Fig. 2. 시스템 개념도

그림 2는 본 연구에서 제안하는 시스템의 개념도이다. 사용자가 SNS를 통하여 해시태그 검색을 한 후 정렬된 화면을 보여주고, 결과 페이지의 데이터를 자동으로 수집하기 위해 크롤링을 진행한다. 크롤링을 통해 수집한 데이터는 KoNLPY(코엔엘파이, 한국어 형태소 분석기)를 사용하여 각 본문의 키워드(명사)를 추출한다. Cosine Similarity(코사인 유사도, 유사도 분석)를 사용하여 두 벡터(해시태그 검색어와 추출한 본문 키워드) 간의 코사인 각도를 검사해 결괏값을 도출한다. 이에 따라 나온 결괏값을 통해 검색 결과 페이지 노출 순서를 재정렬한 후 사용자에게 해시태그 검색어 재정렬 화면을 제공한다.

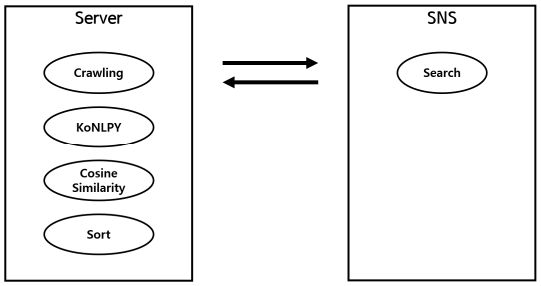

Fig. 3. 시스템 구성도

그림 3은 본 연구에서 제안하는 시스템의 구성도이다. SNS(소셜 네트워크 서비스, Social Network Service)의 Search(검색) 모듈을 사용하여 해시태그 검색어를 입력하고, 검색어를 통해 나온 결과 페이지를 보여준다. Server는 Crawling(크롤링) 모듈을 통해 데이터를 자동으로 수집한다. 자동으로 수집한 데이터로 KoNLPY(코엔엘파이) 모듈을 통해 Crawling(크롤링) 한 본문의 키워드(명사)를 추출하고, Cosine Similarity(코사인 유사도) 모듈을 통해 해시태그 검색어와 본문의 키워드(명사)의 유사도를 분석하여 결괏값을 도출한다. 마지막으로 도출된 Cosine Similarity(코사인 유사도) 모듈의 분석 결괏값으로 Sort 모듈을 통해 검색 결과의 재정렬을 진행한다.

IV. 결론

현재 정보기술의 발전으로 인한 소셜 네트워크 서비스(SNS, Social Network Service)는 다양한 형태로 변화하며 지속적으로 발전하고 있다. 특히, 대표적인 SNS의 인스타그램에서의 해시태그(Hashtag) 이용은 사용자들의 검색 활동에 도움이 되고 있다. 해시태그 검색 활동은 도움이 되는 부분도 있는 반면에 이 해시태그 검색을 악용하여 광고성 게시물을 작성함으로써 해시태그 검색 활동의 정확도와 효율성을 감소시키는 문제점이 있다. 본 논문에서는 해시태그 검색어와 본문 내용의 키워드의 유사도 검사를 통하여 검색 결과를 재정렬하는 시스템을 제안하며, SNS 사용자가 해시태그 검색을 통해 높은 정확성과 효율성의 검색 결과를 얻을 수 있도록 한다.

ACKNOWLEDGEMENT

본 연구는 2022년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2019-0-01817)

REFERENCES

[1] 이장호, 윤성로, “짧은 문서에 대한 키워드 추출 알고리즘 성능 향상을 위한 새로운 방법”, 한국정보과학회 동계학술발표회 논문집, 한국정보과학회, pp.578-580, 2015.

[2] 정상원, 정기창, “문자열 유사도 알고리즘을 이용한 공종명 인식의 자연어처리 연구”, 한국건설관리학회 논문집, 한국건설관리학회, vol. 21, No.6, pp.125-134, 2020.

[3] 김상모, 김형준, 한인규, “코사인 유사도 기법을 이용한 뉴스 추천 시스템”, 제25회 한글 및 한국어 정보처리 학술대회 논문집, 한국정보과학회 언어공학연구회, pp.163-166, 2013.