http://dx.doi.org/10.7840/kics.2015.40.12.2485

키워드 추출 및 유사도 평가를 통한 태그 검색 시스템

정 재 인*, 유 명 식°

Tag Search System Using the Keyword Extraction and Similarity Evaluation

Jaein Jung*, Myungsik Yoo°

요 약

해시태그는 현재 페이스북, 트위터와 같은 SNS와 개인 블로그 등에서 활발하게 사용되고 있다. 하지만 스팸성 목적 또는 게시글 조회수 증가 등의 목적으로 무분별하게 해시태그를 사용하여 태그검색의 효율성이 떨어지고 있다. 이에 따라 본 논문에서는 태그검색의 정확도를 높이고자 기존의 키워드 추출 알고리즘과 단어간 유사도 평가 알고리즘을 이용한 태그 검색 시스템을 제안하였다. 제안하는 시스템의 테스트 결과 태그 검색의 정확도가 향상됨을 알 수 있었다.

Key Words : Hashtag, Keyword, Similarity, Search System, mining

ABSTRACT

Recently, Hashtag is widely used in SNS like Facebook, Twitter and personal blogs. However, the efficiency of tag search system is poor due to the indiscriminate use of hashtags. To enhance the accuracy of tag search system, we proposed a tag search system using the keyword extraction and similarity evaluation. The experimental results show that the proposed system provides the higher accuracy on tag search results.

Ⅰ. 서 론

최근 SNS의 급속한 성장으로 인해 SNS 사용자가 작성한 글이 폭발적으로 증가하고, 이러한 환경에서 작성글의 검색 편의성을 높이기 위해 2007년 Stowe Boyd가 해시태그를 제안했다. 해시태그는 #(샤프기호) 뒤에 단어를 작성한 것을 말하며, 게시글을 작성한 후 '#홍길동', '#대한민국' 등의 형태로 게시글 작성자가 직접 작성하여 글의 주제를 나타낸다[1]. 이러한 태그검색을 통해 기존 검색의 특징인 정보 중심의 검색과 SNS의 특징인 관심사 기반의 정보 검색이 가능하도록 하고 있다.

하지만 현재의 태그검색은 해시태그어와 검색어와의 일치성, 작성자의 인기도, 최신성 등을 평가하여 검색 및 결과 노출 순서를 결정하기 때문에 작성한 글과 관련이 없는 해시태그어를 무분별하게 사용하여 자신이 작성한 게시글의 조회수를 높이고 스팸성 목적을 달성한다는 문제점이 있다. 이에 본 논문에서는 이러한 문제점을 개선하기 위하여 기존의 키워드 추출 및 유사도 평가 알고리즘을 통한 태그 검색 시스템을 제안하고자 한다.

Ⅱ. 제안 태그 검색 시스템

앞서 서론에서 언급한 문제점과 사용자의 편의를 위해 해시태그 자동추출이 활발하게 연구된 편이다. 그 중 대표적 방법은 해시태그 자동추출을 위해 TF-IDF(Term Frequency-Inverse Document Frequency)를 적용하여 작성된 글에서의 키워드를 추출하고 그 중 n개를 해시태그어로 추출하는 방법이며, 키워드 추출 방법에 따라서 여러 가지 관련연구가 존재한다[2,3]. 하지만 키워드 추출 시 작성한 글과 연관성이 떨어지는 단어가 키워드로 설정되거나, 글 작성자의 만족도가 떨어지는 단어가 키워드로 설정되는 문제점 때문에 실제로 널리 사용되지는 않고 있다. 그리고 해시태그어간 유사도 평가를 실시하여 유사도가 떨어지는 태그어를 스팸성 태그어로 추출하는 방법도 있으나, 정상적 태그어도 스팸성 태그어로 추출하는 경우 때문에 비효율적이다.

이러한 문제점을 해결하기 위해서 본 논문에서는

※ 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업의 연구결과로 수행되었음 (IITP-2015-H8501-15-1008)
* First Author : Soongsil University, School of Electronic Engineering, wjdwodls111@hanmail.net, 학생회원
° Corresponding Author : Soongsil University, School of Electronic Engineering, myoo@ssu.ac.kr, 종신회원
논문번호 : KICS2015-12-386, Received December 7, 2015; Revised December 15, 2015; Accepted December 15, 2015

'크롤링 된 데이터 내 텍스트와 해시태그어 추출', '텍스트에서의 핵심 키워드 추출', '유사도 계산 및 검색 결과 재정렬의' 3부분으로 나누어진 태그검색 시스템을 제안한다. '텍스트에서의 핵심 키워드 추출'과 '유사도 계산'은 각 페이지마다 각각 시행되며, 그 이후 각 페이지별 유사도 결과값을 내림차순으로 정렬하게 되며, 상세 내용은 다음과 같다.

2.1 텍스트와 해시태그어 추출

첫 번째로, 아래 그림과 같이 태그 검색어 입력 후 나온 많은 결과페이지에서 데이터를 자동으로 수집하기 위해 Crawler를 사용한다. Crawler를 통해 얻은 데이터는 텍스트와 해시태그어로 나누어 따로 저장해둔다. 이 때 Crawler는 Tracking (대상지정 크롤링) 모드 사용을 통해 특정 도메인의 서브도메인 관련 페이지만 수집하여 배너광고 등의 수집을 막아 Crawler의 효율을 높인다.

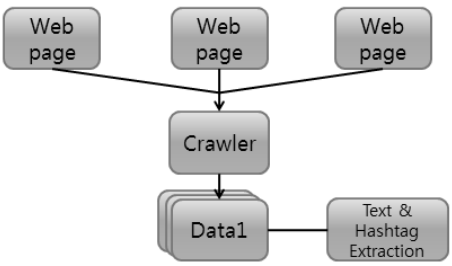

그림 1. 텍스트와 해시태그어 추출
Fig. 1. Extraction of text and hashtag

2.2 텍스트에서의 키워드 추출

두 번째로, Crawler를 사용하여 수집한 데이터 중 따로 저장한 텍스트에 대해서 키워드 추출 알고리즘을 사용하여 각 페이지별 핵심 키워드 추출을 실시한다. 키워드 추출은 텍스트를 형태소 단위로 분할하여 어미와 조사 등의 불용어를 제거하고 난 후 단어의 출현 빈도 파악을 통해 키워드를 추출한다. 이 때 키워드의 출현 빈도가 동일할 경우 다음의 2가지 경우에 따라 가중치를 부여하여 키워드간의 중요도를 파악한다.

- 텍스트의 제목에 해당 단어 존재 시 가중치
- 텍스트 내 특수처리(굵게, 색상) 된 부분에 키워드 존재 시 가중치

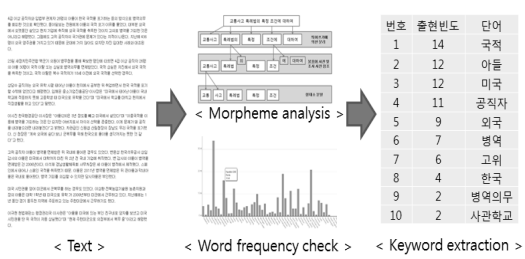

그림 2. 텍스트의 형태소 분석
Fig. 2. Morpheme division of the text

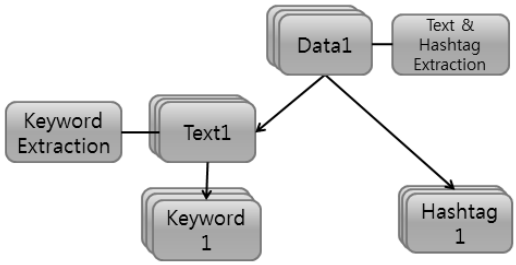

그림 3. 핵심 키워드 추출
Fig. 3. Keyword extraction

2.3 유사도 계산 및 검색 결과 정렬

마지막 세 번째로, 저장해둔 해시태그어와 추출한 키워드간의 유사도를 각 페이지별로 계산하여 페이지별 유사도 점수 결과값을 도출한다. 키워드와 해시태그어 간의 유사도 계산은 두 개의 문자열의 유사도를 계산하는 법으로 널리 알려진 Levenshtein Distance 알고리즘을 사용한다. 이 때 글 작성자가 만든 해시태그어가 n개일 경우 키워드 추출 알고리즘을 통해 추출한 키워드 n개와 유사도를 계산하여 한 페이지의 n개 해시태그어에 대한 유사도 평균을 도출한다. 같은 방식으로 각각의 페이지에 대해 유사도 점수 평균 결과값을 도출하고, 페이지별 결과값에 따른 내림차순 정렬을 통해 검색 결과 페이지 노출 순서를 재정렬하여 검색 결과페이지 상단에 위치할수록 검색어에 대한 정확도가 높은 페이지가 위치하도록 한다.

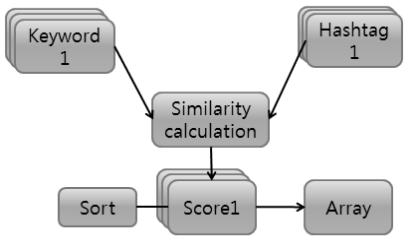

그림 4. 유사도 계산 및 검색 결과 재정렬
Fig. 4. Similarity calculation and sorting

Ⅲ. 제안 시스템 성능 평가

제안하는 시스템의 성능 분석을 위해 표 1과 같이 4개 분야별 인기 검색어 3개씩 총 12개의 검색어를 설정하고 성능 분석을 실시하였으며, 각 검색어 검색 결과 페이지수의 통일을 위해 검색 결과 상위 노출 순서로 50개의 페이지에 대해서 성능 분석을 실시하였다. 또한 결과 그래프의 시각적 편의성을 위해 50개 페이지를 상위 노출 순서로 10개씩 총 5개의 그룹으로 편성하여 그래프로 나타내었다.

대표적으로 IT 분야의 아이폰 태그검색의 성능 테스트 결과는 그림 5와 같다. 그림 5에서 유효 페이지는 아이폰 사용후기, 스펙정보, 사용법 등의 정보를 가진 페이지를 뜻하며, 아이폰 판매처 홍보 등과 같은 스펨성 페이지는 비유효 페이지로 정의하였다. 기존 검색 결과의 경우 상위 50개 페이지 중 검색어와 관련성이 떨어지는 19개 비유효 페이지가 전 구간에 비교적 고르게 분포하는 것을 확인할 수 있었으며, 제안하는 시스템의 테스트 결과 검색어와 관련성이 떨어지는 19개 비유효 페이지의 노출 순서가 기존대비 하위로 바뀌어 상위 노출 된 페이지들의 정확도가 향상 된 것을 확인할 수 있었다.

같은 방식으로 표 1에서 언급한 4개 분야 12개 단어를 테스트한 후 12개 검색어의 검색 결과를 구간별로 총합하여 정확도 평균값을 구하였다. 여기서 정확도는 검색결과 50개 페이지를 상위 노출 순으로 10개씩 편성한 그룹마다 그룹 내 비유효 페이지를 제외한 유효 페이지의 개수로 정의하였다. 테스트 결과는 표2와 같으며, 기존 결과와 비교하여 검색결과 상단에 위치하는 페이지의 정확도가 증가하고 검색결과 하단 쪽에는 검색어와 관련성이 낮은 페이지를 위치시킴에 따라 정확도가 낮아짐을 확인할 수 있었으며, 이를 통해 검색결과의 정확도와 효율성이 높아짐을 확인할 수 있었다.

표 1. 4개 분야 검색어 설정
Table 1. Search word set in 4 areas

Field	Word	Field	Word
IT	Iphone	Sports	Yu-na Kim
	Galaxy		Soccer
	Tablet PC		Entertainer
News	Ebola	Others	Soongsil Univ.
	IS		Winter
	North korea		Christmas

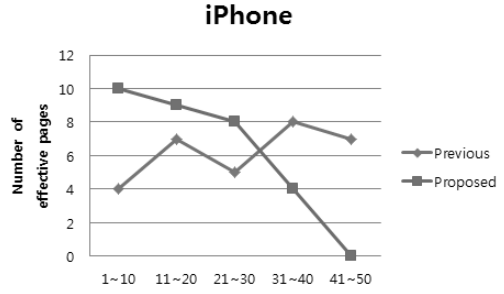

그림 5. 아이폰 태그검색 테스트 결과
Fig. 5. Search results of iphone tag

표 2. 태그 검색 결과
Table 2. Tag search results

Sequence of search results	Accuracy of previous system (%)	Accuracy of proposed system (%)
1~10	73.3	90.3
11~20	63.3	85.0
21~30	60.8	73.3
31~40	68.3	55.8
41~50	57.5	16.6

Ⅳ. 결 론

최근 SNS는 물론 개인 페이지 운영자들의 해시태그에 대한 관심과 사용이 증대되고 있다. 본 논문에서는 기존의 키워드 추출 알고리즘과 단어간 유사도 평가 알고리즘을 사용하여 태그 검색의 정확도와 효율성을 높이고자 하였으며, 모의실험을 통하여 제안 시스템의 타당성 검증을 수행하였다.

References

[1] S. Oh, “Marketing, indulge in a sea of Hashtags(#),” *Marketing 2015*, vol. 49, no. 10, 59-64, Oct. 2015.

[2] S. M. Kywe, T.-A. Hoang, E.-P. Lim, and F. Zhu, “On recommending hashtags in twitter networks,” *Social Informatics,* Springer Berlin Heidelberg, vol. 7710, pp. 337-350, 2012.

[3] E. Zangerle, W. Gassler, and G. Specht, “Recommending#-tags in twitter,” in *CEUR Workshop Proc. SASWeb 2011*, vol. 730, pp. 67-78, 2011.