JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 19, No. 1, pp. 157-163, Jan. 2018

트위터에서 형태소 분석과 PageRank 기반 화제단어 추출 방법 제안

이 원 형 · 조 성 일 · 김 동 회*
강원대학교 IT대학 전기전자공학부

Proposal of keyword extraction method based on morphological analysis and PageRank in Tweeter

Won-Hyung Lee · Sung-Il Cho· Dong-Hoi Kim*
Kangwon National University, IT College, Electronic Electronics Engineering

[요 약]

SNS를 이용하는 사람들은 매일 자신의 다양한 생각을 SNS에 게시한다. SNS에 게시된 데이터는 수많은 사람들의 생각과 의견이 담겨있다고 할 수 있다. 특히 트위터에서 서비스되는 인기 화제어는 사용자가 올린 글에서 자주 등장한 단어의 횟수를 집계해 순위를 결정한다. 하지만 이와 같은 방법은 단순히 중복된 단어가 나열된 불필요한 데이터에 민감하다. 제안하는 방법은 단어간의 관계도를 이용한 단어의 화제성을 기반으로 순위를 결정하므로 불필요한 데이터의 영향을 적게 받고 주요단어를 안정적으로 추출할 수 있다. 성능 비교를 위하여 내림차순 화제어 순위와 상위 20개중에서 의미 없는 화제어의 비율 측면에서 형태소 분석과 PageRank 기반의 제안 방식과 단순 등장 횟수 기반의 기존 방식을 비교한다. 제안하는 방안과 기존 방안은 상위 20개중에서 무의미한 화제어를 각각 55%과 70%를 순위권에 포함시켰으며 제안한 방법이 기존 방법과 비교할 때 15% 정도 향상된다.

[Abstract]

People who use SNS publish their diverse ideas on SNS every day. The data posted on the SNS contains many people's thoughts and opinions. In particular, popular keywords served on Twitter compile the number of frequently appearing words in user posts and rank them. However, this method is sensitive to unnecessary data simply by listing duplicate words. The proposed method determines the ranking based on the topic of the word using the relationship diagram between words, so that the influence of unnecessary data is less and the main word can be stably extracted. For the performance comparison in terms of the descending keyword rank and the ratios of meaningless keywords among high rank 20 keywords, we make a comparison between the proposed scheme which is based on morphological analysis and PageRank, and the existing scheme which is based on the number of appearances. As a result, the proposed scheme and the existing scheme have included 55% and 70% of meaningless keywords among high rank 20 keywords, respectively, where the proposed scheme is improved about 15% compared with the existing scheme.

색인어 : 트위터, 화제성 높은 단어, 페이지랭크 알고리즘, 형태소 분석
Key word : Twitter, Trend keyword, PageRank algorithm, Morphological analysis

http://dx.doi.org/10.9728/dcs.2018.19.1.157

Received 02 November 2017; **Revised** 23 January 2018
Accepted 29 January 2018

***Corresponding Author; Dong-Hoi Kim**

Tel: +82-033-250-6349
E-mail: donghk@kangwon.ac.kr

Ⅰ. 서론

사회는 인터넷을 통해 빠르게 변화하고 있다. 2014년 1분기 실적에 따르면 전 세계적으로 SNS이용자수가 12억7천만 명을 넘어섰고, 일 이용자도 8억 명을 돌파하였다. SNS 서비스 중 하나인 트위터 역시 월간 활동이용자 수가 1천 3백만 명에 이른다[1]. SNS 사용자들은 SNS를 이용해 자신의 의견을 표출하고 대화를 한다. 이러한 규모가 큰 SNS의 사용은 대선 등의 선거에도 영향을 미쳐서 결과를 예측하거나 유력 후보 등을 파악할 수 있다[2]. 또한 SNS는 고객와 기업, 국민과 정부 사이의 대화의 창구가 되기도 한다. SNS 사용자들이 나눈 방대한 대화기록은 각 시대상을 대표하고 있는 주요 단어를 내포하고 있다. 시대상을 대표하고 있는 주요 단어를 활용하기 위해 많은 포털사이트들이 검색 횟수를 집계해 화제가 되는 단어순위를 만든다[3]. 하지만 이는 단순히 검색 횟수만을 집계해 순위를 만들기에, 악의를 가진 사용자들이 반복적으로 한 단어를 검색하는 등의 행동에 순위가 영향을 많이 받는다. 본 논문은 단순히 검색 횟수나 등장 횟수가 아닌 단어간의 관계도를 그린 뒤 관계도를 기반으로 가중치를 계산한 결과로 화제가 되는 단어순위를 만드는 방법을 제안한다.

중요한 단어를 찾는 비슷한 기존의 알고리즘으로는 TextRank 알고리즘이 존재한다[4]. 하지만 이 TextRank 알고리즘은 한 가지 주제의 글에서 주제어가 될 가장 중요한 단어를 추출하는 알고리즘이고, 본 논문이 제안하는 알고리즘은 여러 다양한 주제의 문장의 집합에서 가장 많이 언급된, 가장 화제가 되는 단어를 추출하는 알고리즘이다. 알고리즘의 추구 목적이 달라서 성능 비교하기가 사실상 어렵기 때문에 비교 대상으로 사용하지 않았다.

본 논문의 Ⅱ장에서는 SNS중 하나인 트위터의 특성과 이를 활용하는 방안에 대해 서술하였다. Ⅲ장에서는 기존 방안을 확장하여 웹 데이터를 수집하는 방법과 저장된 데이터를 분석하는 방법에 대해 서술하였다. Ⅳ장에서는 실험 방법과 실험 결과를 분석하였고, 마지막으로 Ⅴ장에서는 본 논문의 결론을 제시한다.

Ⅱ. 기존의 등장 횟수에 따른 화제 단어 파악 방법

트위터는 사회 관계망 서비스 중 하나로, 여러 사용자가 자기 생각이나 현재 상황 등을 짧은 글로 표현할 수 있는 플랫폼이다. 수많은 사람이 트위터에 자신의 생각이나 관심거리를 업로드하게 되고, 이렇게 생산된 방대한 양의 데이터는 사용자의 생각과 관심거리를 내포하고 있다. 이 데이터를 잘 분석하면 데이터가 업로드된 시대의 추세와 화제를 파악할 수 있다. 데이터의 이러한 특성을 이용하기 위한 기존 방법은 주로 등장 횟수나 비율을 집계해 화제성 높은 단어순위를 결정하는 방식이다. 이는 주로 포털사이트 등에서 사용하며, 가장 많이 사용하는 방법이다. 하지만 이런 방법을 사용한 순위는 사용자가 사용한 특정한 단어를 기반으로 하기 때문에 등장 단어가 편협하고 악의를 갖고 특정 단어를 반복해 검색하는 등 악영향을 받았기 때문에 신뢰하기 힘든 데이터이다[5]. 트위터에서도 위와 같은 방법을 사용하며, 사용자의 편협한 단어 사용범위와 홍보를 위한 특정 단어 반복 사용에 취약하다.

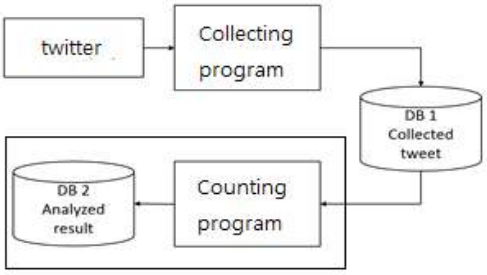


그림 1. 단순 등장 횟수 기반의 기존 방법을 사용한 프로그램 구조의 예시
Figure 1. Example of program structure using the existing method based on the number of appearances

그림 1은 기존방법을 사용하여 분석하는 프로그램의 예시다. Collecting program은 트위터에서 사용자가 올린 문장을 수집해 DB1에 해당 문장이 올라온 년, 월과 함께 저장한다. Counting program은 DB1에서 문장을 하나씩 가져와 단어로 분할한 후 단어의 등장회수를 세어 DB2에 저장한다. 제안하는 방법은 박스로 감싸여진 부분을 수정했다.

Ⅲ. 제안하는 형태소 분석과 PageRank 알고리즘을 이용한 화제 단어 파악 방법

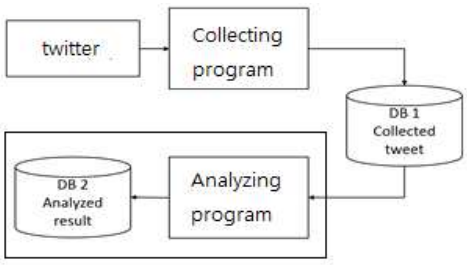


그림 2. 형태소 분석과 PageRank 기반의 제안하는 방법을 사용한 프로그램 구조의 예시
Figure 2. Example of program structure using the proposed method based on morphological analysis and PageRank

그림 2는 형태소 분석과 PageRank 기반의 제안하는 분석 방법을 사용한 프로그램의 예시다. Collecting program과 DB1은

동일하게 사용하며 Analyzing program과 DB2에 제안하는 분석 방법을 적용하여 사용하였다. 박스로 감싸여진 부분을 상세하게 나타내면 다음과 같다.

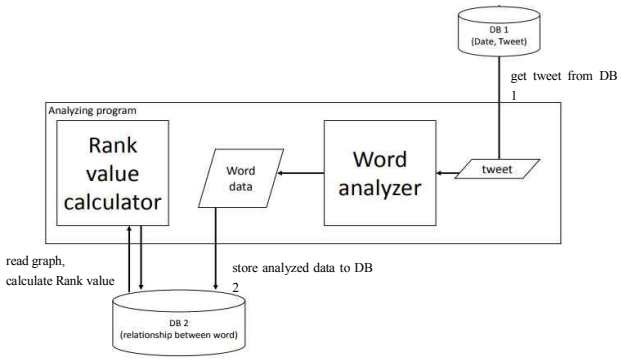


그림 3. 제안하는 분석 방식의 자세한 프로그램 구조
Figure 3. Detailed program structure of proposed analysis method

그림 3은 제안하는 분석 방식의 자세한 프로그램 구조를 보여 주고 있다. 그림 3에서 DB 1은 원본 데이터, DB 2 는 제안하는 방식을 저장할 공간이다. 그림 3에서 보는 바와 같이 제안하는 방식은 3단계로 이루어진다. 먼저 문장을 단어로 분할하는 형태소 분석, 분할된 단어를 관계도의 형태로 저장하는 DB 2, 마지막으로 생성된 단어간 관계도를 통해 가중치를 계산하는 PageRank 알고리즘이다. 그림 3을 자세히 설명하면 다음과 같다. 먼저 Word Analyzer는 원본 데이터를 수집해 저장한 DB 1에서 문장을 꺼내 문장을 형태소별로 나눈 뒤 단어끼리 연결한다. 연결한 결과를 DB 2에 저장되어있는 그래프에 추가한 뒤 저장한다. Rank Value Calculator는 DB 2에 만들어진 그래프를 이용해 PageRank 알고리즘을 이용하여 단어별 가중치를 계산한 후 다시 넣어준다.

첫 번째 단계로 그림 3에서 Word analyzer 로 표현하고 있으며 문장을 단어로 분할한다. 그러나 한국어의 특성상 용언의 활용이 다양하므로, 단순히 띄어쓰기를 기준으로 단어로 나누게 되면 단어의 변형, 띄어쓰기 오류 등을 구분해 낼 수 없다. 그래서 형태소 분석을 통해 문장을 비교적 정확하게 단어로 분할해 준다. 형태소 분석은 문장을 먼저 자소단위로 분할 한 다음에 이를 조합해 나가며 가장 가능성이 높은 단어를 분별한다. 일반 명사와 고유명사만 사용한다. 예를 들어 '감기는'이란 문장을 구분하고자 한다.

그림 4는 자소단위로 분할된 단어를 조합하기 위한 Trie구조의 예시다. '감기는' 이란 문장을 자소단위로 구분하면 ㄱ, ㅏ, ㅁ, ㄱ, ㅣ, ㄴ, ㅡ, ㄴ 으로 나누어진다. 이를 순서대로 조합해 보면 '감', '기', '는' 으로 조합 될 수도 있고, '감기', '는' 으로 조합 될 수도 있다. 여러 가지 상황 중 가장 가능성이 높은 '감기', '는' 으로 구분되어 분할시킨다.

두 번째 단계로 그림 3에 있는 Word analyzer를 통해 분해된 Word data를 이용해 관계도를 생성한다. 수집하는 데이터의 특성상 모두 짧은 문장이기 때문에 각 단어는 포함된 문장 내의 모든 단어와 연관성이 있다고 보고 각각 관계를 이어주며 모두 양방향으로 생성된다.

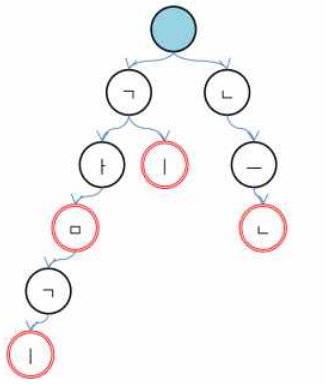


그림 4. 자소 단위 Trie 구조의 예[6]
Figure 4. Truncated trie structure

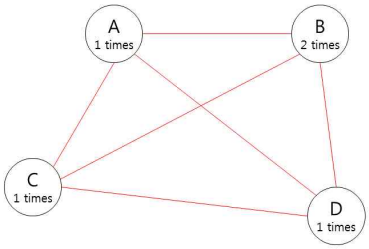


그림 5. A-B-C-D-B 형태의 문장을 분석한 그래프
Figure 5. Graph from analysis of sentences in the form of A-B-C-D-E

그림 5는 A-B-C-D-B 형태의 문장을 분석한 그래프이다. 단어 B의 경우 두 번 등장하였다. 이때 한 문장 내에서 중복된 단어는 노드를 새로 추가하지 않고 기존 단어에 등장 횟수를 추가시켜 제안하는 알고리즘과 비교할 등장 횟수를 집계한다. 그림 5에 새로운 문장을 추가시킬 경우 마찬가지로 중복 단어는 새로 추가하지 않고 서로 공유하며 관계도를 확장시켜 나간다.

그림 6은 그림 5에 B-C-D-E-F 문장이 추가된 그래프이다. 기존 링크는 검정색, 추가된 링크는 붉은색으로 표시하였다. 단어 B, C, D가 먼저 생성된 단어와 중복된다. 중복단어는 등장 횟수를 추가시키고, 새로운 단어는 기존 그래프에 계속해 추가한다. 단어간 관계를 나타내는 간선은 중복된 수를 사용하지 않기 때문에 따로 횟수를 기록하거나 새로운 간선을 이어주지 않는다. 완성된 그래프에 제안하는 방법을 사용하여 단어별 가중치를 계산한다.

마지막 단계로 단어의 가중치(중요도)를 구하기 위해 그래프 기반 가중치 알고리즘인 PageRank 알고리즘을 응용하여 그림 3에서 Rank value calculator에서 수행한다[7]. 그래프의 한 노드인 A의 가중치인 PR(A)은 다음의 식을 이용해 구한다.

$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}) \quad (1)$$

여기서 d는 그래프의 각 노드에서 링크를 타고 넘어갈 확률이며, 이 값은 원 논문과 같은 0.85로 고정한다. C는 해당 노드가 갖고 있는 링크의 수이다. 단어 A의 가중치를 구하기 위해서는 인근 노드의 가중치를 링크의 수로 나눈 값을 더해나가는 과정이 필요하다. 이는 인근 노드로부터 가중치를 골고루 나누어 받는 과정이라고 할 수 있다. 이 과정을 반복하면 해당 단어의 가중치에 수렴한다. PageRank 알고리즘을 단어에 적용하기 위해 원 알고리즘의 d 값은 한 문장에서 해당 단어가 다른 단어들과 연관이 있을 확률, C 값은 단어가 가지고 있는 연관된 단어들의 수로 사용한다. 이 때, 단어의 등장 횟수는 가중치에 영향을 주지 않는다.

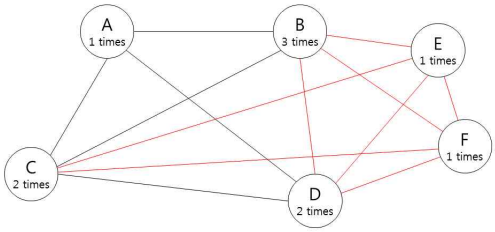


그림 6. 그림 5에 B-C-D-E-F 문장이 추가된 그래프
Figure 6. Graph with B-C-D-E-F sentence added in Figure 5

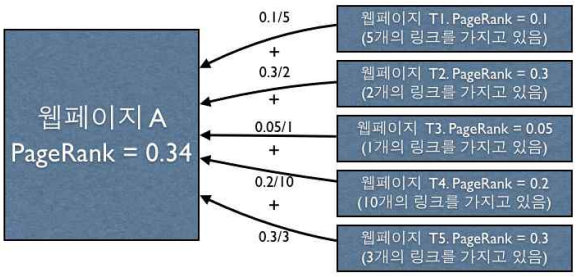


그림 7. 가중치를 구하는 예시[6]
Figure 7. example of calculating rank value

그림 7은 PageRank 알고리즘에서 가중치를 구하는 예시를 그림으로 설명한 것이다. 웹페이지 A의 가중치를 구하기 위해서 주변 웹페이지가 자신의 링크의 수만큼 골고루 가중치를 나누어주는 것을 볼 수 있다. 웹페이지가 갖고 있는 링크가 많을수록 주변으로부터 많이 받고 적게 나눠주는 것을 알 수 있다.

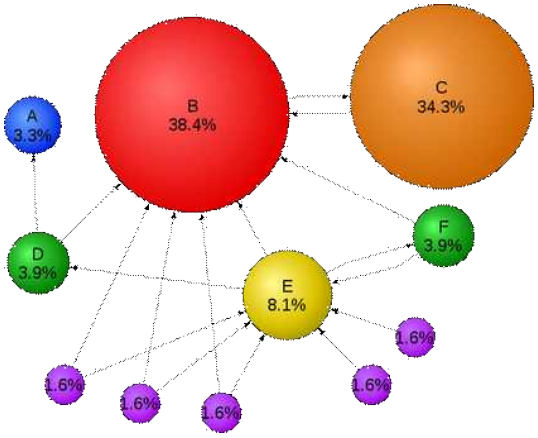


그림 8. PageRank 가중치의 예시[6]
Figure 8. Example of PageRank weights

그림 8은 가중치가 무엇을 기준으로 높아지는지를 보여준다. B를 보면 많은 노드들이 가중치를 주고 있기에 높은 가중치를 가지고 있다. 가장 큰 특징은 C의 가중치다. B하고만 연결이 되어있지만 B가 유일하게 가중치를 나누어 주는 대상이기 때문에 높은 가중치를 갖고 있다. 이를 본 논문의 단어와 연관지으면 두 단어가 같이 쓰이는 경우 유사한 가중치를 갖게 되어 분석이 용이해진다. 본 논문의 모든 단어간 관계는 양방향으로 이어져 있기에 그림 7과 같은 극단적으로 중요도가 치우치지 않는다. 즉 단어간 관계도에 PageRank 알고리즘을 적용하면 화제가 되는 단어일수록 다양한 문장에 등장할 것이고 자연스럽게 가중치가 올라간다.

Ⅳ. 실험환경 및 분석결과

제안하는 방식의 효율성을 확인하기 위해서는 많은 양의 데이터를 필요로 한다. 따라서 웹상을 돌아다니며 데이터를 수집하는 프로그램을 별도로 만들어 이용하였다. Collecting program은 트위터의 계정들을 거미줄처럼 퍼져 나가며 데이터를 수집한다. 이는 해당 계정의 팔로우, 팔로워 계정으로부터 1달 이내의 데이터를 가져와 DB 1에 저장하는 역할을 한다. 데이터가 수집된 계정은 DB 1에 따로 저장되어 중복된 계정으로부터 또다시 데이터를 수집하지 않는다. 여기에는 총 5000개의 계정이 저장되며 5001번째로 수집되는 계정은 5000번째로 들어가게 되고 1번에 있던 계정은 다시 방문이 될 경우 수집된다. 즉, 위로 한 칸씩 밀며 계정을 수집한다. 데이터를 수집하기 위해 Collecting program은 50개의 계정을 저장할 수 있는 큐에 해당 계정의 팔로우, 팔로워 계정 중 10개의 계정을 무작위로 선별하여 넣는다. 그 후 Collecting program은 큐의 최상위 계정으로 이동하게 된다. 큐는 중복을 허용하지 않으며 따라서 50개의 계정을 돌아다닐 동안은 중복된 계정에 들르게 되는 일은 없다. 경기도 공식 트위터를 기준으로 데이터를 수집해 퍼져나간다.

표 1. 환경
Table. 1. Environment

Environment	Tool
Word extractor	KOMORAN
Language	Java
Database 1	MySQL
Database 2	Neo4j
program editor	Eclipse

표 1은 데이터 분석환경이다. 형태소 분석을 위해 오픈소스 라이브러리 KOMORAN[8]을 사용하였다. DB 1은 MySQL로 [9], DB 2는 Neo4j[10]로 만들어졌다.

```
mysql> SELECT table_name, table_rows, round(data_length/(1024*1024),2) as 'DATA_
SIZE(MB)', round(index_length/(1024*1024),2) as 'INDEX_SIZE(MB)' FROM informatio
n_schema.TABLES where table_schema = 'joljak' GROUP BY table_name ORDER BY data_
length DESC LIMIT 10;
+------------+------------+---------------+----------------+
| table_name | table_rows | DATA_SIZE(MB) | INDEX_SIZE(MB) |
+------------+------------+---------------+----------------+
| tweet      |      24637 |          4.52 |           0.00 |
| worked     |       2625 |          0.16 |           0.00 |
+------------+------------+---------------+----------------+
2 rows in set (0.00 sec)

mysql>
```

그림 9. 수집한 문장의 개수와 방문한 계정의 개수
Figure 9. collected sentences and visited accounts

그림 9는 수집한 문장의 개수와 방문한 계정의 개수를 확인한 결과이다. 성능의 분석을 위해 24637개의 문장 데이터를 수집해 분석하였다. 14418종류의 단어로 분할되었고 단어간 관계는 363249개 이다.

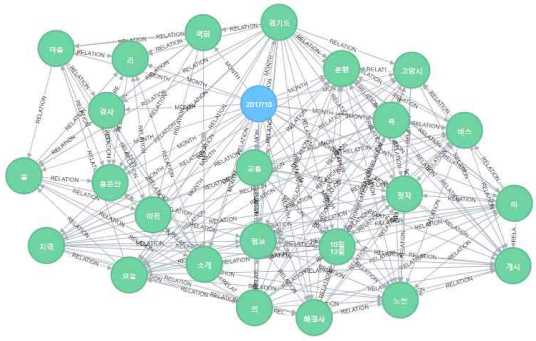


그림 10. 24637개의 트윗 샘플을 이용해 그린 그래프
Figure 10. graph generated with 24637tweet samples

그림 10은 수집한 데이터를 이용해 그린 그래프의 일부이다. 초록색은 단어 노드이며, 각 노드마다 Keyword, Count, Rank 값이 저장되어있다. 2017/10라 적힌 파란색은 날짜 노드이며, 단어가 등장한 날의 모든 노드와 연결된다. 이를 활용해 전체 그래프의 분석 수치, 각 월별 분석 수치를 각각 구하게 된다. 다음은 기존의 방법인 등장 횟수에 따른 내림차순으로 정렬한 화제어 순위와 가중치에 따라 내림차순으로 정렬한 화제어 순위이다.

표 2. 기존의 방식에의한 내림차순 화제어 순위
Table 2. The descending keyword rank by the existing scheme number of appearances

Keyword	Rank	Count
사람	42.618408203125	685
신천지	10.138364791870117	682
대통령	31.953882217407227	660
안마	4.387841701507568	642
때	45.6072883605957	593
말	37.78461837768555	512
비교	8.743020057678223	486
오늘	35.65842056274414	477
교리	7.223877429962158	448
박근혜	24.001638412475586	437
출처	27.085432052612305	409
감사	13.412151336669922	405
편지	8.112653732299805	376
평화	10.101985931396484	374
문재인	25.025537490844727	371
안	28.44438934326172	334
사랑	18.446746826171875	324
뉴스	24.246732711791992	321
세월호	12.717619895935059	320
구속	16.640819549560547	310

표 3. 제안된 방식에 의한 내림차순 화제어 순위
Table 3. The descending keyword rank by the proposed scheme

Keyword	Rank	Count
때	45.6072883605957	593
사람	42.618408203125	685
말	37.78461837768555	512
오늘	35.65842056274414	477
대통령	31.953882217407227	660
안	28.44438934326172	334
출처	27.085432052612305	409
문재인	25.025537490844727	371
뉴스	24.246732711791992	321
국민	24.1080322265625	290
박근혜	24.001638412475586	437
의원	23.553447723388672	289
일	23.54977798461914	296
후	23.25676918029785	189
시간	22.485368728637695	285
정부	21.742490768432617	253
생각	21.70583724975586	285
전	21.62595558166504	186
한국	21.533607482910156	200
시작	19.492883682250977	165

표 2와 표 3은 각각 기존 방법인 등장 횟수를 이용한 순위와 제안하는 방식인 단어간 관계도를 통한 가중치를 사용한 순위를 20위까지 나타냈다. 공통적으로 '때', '말', '출처', '안'이라는 단어가 순위권에 올랐다. 이 단어들은 다양한 문장형에 두루 쓰이는 단어이기 때문에 특별한 화제성이 없음에도 불구하고 높은 순위권에 올랐다. 특히 '안'이라는 단어는 부정의 의미로

‘안 했다’등으로 쓰이는 부사이지만, 일반적으로 ‘안했어’처럼 붙여서 사용하기 때문에 형태소 분석과정에서 명사로 분류되었다. 등장 횟수를 사용한 표 2의 경우 ‘신천지’, ‘교리’, ‘비교’, ‘편지’, ‘안마’ 등의 단어가 순위권에 올랐지만 제안한 방법을 사용한 표 3에는 없었다.

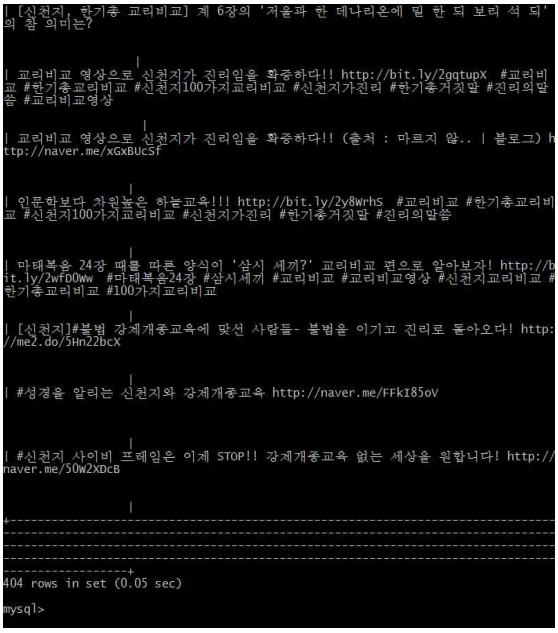


그림 11. 원본데이터 중 신천지 검색 결과
Figure 11. Shinchonji search result of the original data

그림 11은 원본 데이터 중 신천지를 조회 한 결과를 보여준다. ‘신천지’, ‘교리’, ‘비교’, ‘편지’는 수집 과정 중 특정 종교의 여러 계정에서 홍보하는 글을 수집하였기 때문에 높은 순위를 가지게 되었다. 비슷한 이유로 성매매를 뜻하는 은어인 ‘안마’를 사용한 광고를 수집하였기 때문이다. 반대로 제안하는 방법인 표 3은 ‘전’, ‘후’, ‘일’ 등의 단어가 많았으나 등장 횟수를 사용한 표 2에는 없었다. ‘때’, ‘말’, ‘안’ 단어가 순위권에 오른 이유와 같지만 단순하게 등장 횟수가 적었기 때문이다. 즉 제안하는 방법은 등장 횟수가 영향을 끼치지 않음을 확인할 수 있다.

그림 12는 제안한 방식과 기존 방식에서 각각 상위 20개 화제어 중 무의미한 화제어와 시기와 관계없이 두루 쓰이는 화제어들의 수를 나타낸 것이다. 무의미한 화제어는 구하고자 하는 화제어와 반대되는 화제어들이므로 높을수록 낮은 효율을 의미한다. 무의미한 화제어는 광고 등 악의적인 의도로 사용된 것으로 추정된 단어 또는 다른 단어와 함께 사용되어 의미를 갖는 단어를 의미한다. 트렌드를 반영하지 않는 무의미한 화제어는 가중치로 정렬한 순위에서 11개, 등장 횟수로 정렬한 순위에서 14개로 확인되었다. 또한 순위가 높아질수록 등장 횟수로 정렬한 순위의 무의미한 화제어 비율이 높아진다. 이는 제안하는 방식이 55%, 기존 방식이 70%의 무의미한 단어를 보이며 15%의 효율 상승을 나타낸다.

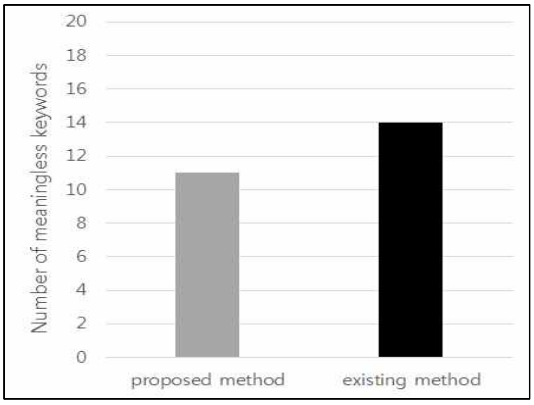


그림 12. 제안된 방식과 기존방식에서 상위 20개 화제어 중에서 무의미한 화제어의 수
Figure 12. The number of meaningless keywords among high rank 20 keywords in the proposed scheme and existing scheme

Ⅴ. 결 론

본 논문에서는 트위터를 활용해 화제성이 높은 단어를 도출하는 과정으로 웹상에서 돌아다니는 프로그램과 이를 가져와 단어로 분할하고 관계도를 생성하여 가중치로 나타낸 결과를 보였다. 기존 검색어 순위 등에서 사용하는 등장 횟수를 사용한 화제 단어는 광고계정, 특정 집단의 반복 데이터 업로드 등에서 취약한 모습을 보였다. 반면 본 논문이 제안하는 가중치를 사용한 방법은 광고, 반복 데이터 업로드와 같은 편향된 정보 속에서도 안정적으로 화제성이 높은 단어를 도출해내는 결과를 보였다. 향후 순위에 오른 단어들에서 광범위하게 사용되지만 특별한 의미가 없는 단어를 제외할 방안의 연구가 필요하다. 추후 프로그램을 확장시켜 문장의 구조, 해당 단어를 사용할 때 함께 사용되는 단어 등이 관계도의 형태로 저장되어 있는 만큼 문장의 자동완성, 추천 단어 리스트 등의 부가 기능을 넣는 것으로 응용할 수 있다.

감사의 글

2017년도 강원대학교 대학회계 학술연구조성비로 연구하였음 (관리번호-520170064)

참고문헌

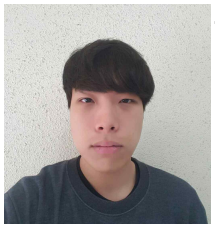
[1] Yun-hi Lee. Use of domestic SNS and analysis of major issues. Internet & Security Focus, 2014, 10.
[2] Chang-Jin Han, Kyoung-Soo Kim. “Twitter's impact on the election of TV debates -18th presidential election TV debates”. 2013
[3] Search term ranking [Internet]

http://datalab.naver.com/keyword/realtimeList.naver
[4] Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into text." Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.
[5] ji-Yeon. Search bias issues on portal and effective search values : focusing on keyword searcheso 'Naver', 2016.
[6] 조성문의 블로그, '쉽게 설명한' 구글의 페이지 랭크 알고리즘', Aug 26 2012, https://sungmooncho.com/2012/08/26/pagerank/, Oct 16 2017
[7] PAGE, Lawrence, et al. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.
[8] KOMORAN[Internet]. Available: http://shineware.tistory.com/entry/KOMORAN-30-beta
[9] MySQL[Internet]. Available: https://www.mysql.com/
[10] Neo4j[Internet]. Available: https://neo4j.com/

이원형(Won-Hyung Lee)

2017년～현재: 강원대학교 IT대학 전자통신공학과 학사과정

※관심분야 : S/W 알고리즘 설계 및 응용 프로그램 개발



조성일(Sung-Il Cho)

2017년～현재: 강원대학교 IT대학 전자통신공학과 학사과정

※관심분야 : S/W 알고리즘 설계 및 응용 프로그램 개발

김동회(Dong-Hoi Kim)

2005년 : 고려대학교 전파공학과 (공학박사)

1989년 1월 ~ 1997년 1월 : 삼성전자 전임연구원
2000년 8월 ~ 2005년 8월 : 한국전자통신연구원 선임연구원
2006년 3월 ~ 현재 : 강원대학교 IT대학 전기전자공학부 전자통신학과 교수

※관심분야 : 이동통신 및 무선 네트워크 등