JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 2, pp. 123-131, February 2020
https://doi.org/10.9708/jksci.2020.25.02.123

Performance Evaluations of Text Ranking Algorithms

Myung-Hwi Kim*, Beakcheol Jang*

*Student, Dept. of Computer Science, Sangmyung University, Seoul, Korea
*Professor, Dept. of Computer Science, Sangmyung University, Seoul, Korea

[Abstract]

The text ranking algorithm is a representative method for keyword extraction, and its importance is emphasized highly. In this paper, we compare the performance of recent research and experiments with TF-IDF, SMART, INQUERY and CCA algorithms, which are used in text ranking algorithm.. After explaining each algorithm, we compare the performance of each algorithm based on the data collected from news and Twitter. Experimental results show that all of four algorithms can extract specific words from news data equally. However, in the case of Twitter, CCA has the best performance to extract specific words, and INQUERY shows the worst performance. We also analyze the accuracy of the algorithm through six comparison metrics. The experimental results present that CCA shows the best accuracy in the news data. In case of Twitter, TF-IDF and CCA show similar performance and demonstrate good performance.

▸ **Key words**: Text ranking algorithm, TF-IDF, SMART, INQUERY, CCA

[요 약]

텍스트 순위 알고리즘은 키워드 추출을 위한 대표적인 방법이며 그 중요성이 강조되고 있다. 본 논문에서는 텍스트 랭킹 알고리즘에서 대표적으로 사용되는 TF-IDF, SMART, INQUERY, CCA 알고리즘이 적용된 최근 연구와 실험해비교한다. 먼저, 각 알고리즘을 설명한 후 뉴스와 트위터 데이터를 기반으로 알고리즘의 성능을 분석한다. 실험 결과에 따르면 네 가지 알고리즘 모두 뉴스 데이터에서 특정 단어의 추출 성능이 좋다는 것을 알 수 있다. 그러나 Twitter의 경우 CCA는 특정 단어를 추출하는 최고의 성능을 가지며 INQUERY는 가장 낮은 성능을 보여준다. 또한 6 가지 비교 메트릭을 통해 알고리즘의 정확성을 분석한다. 실험 결과 CCA가 뉴스 데이터에서 최고의 정확도를 보여주고, 트위터의 경우 TF-IDF와 CCA는 비슷한 성능을 보이며 높은 정확도를 보인다.

▸ **주제어**: 텍스트 랭킹 알고리즘, TF-IDF, SMART, INQUERY, CCA

• First Author: Myung-Hwi Kim, Corresponding Author: Beakcheol Jang
*Myung-Hwi Kim (nopkgogo3@gmail.com), Dept. of Computer Science, Sangmyung University
*Beakcheol Jang (bjang@smu.ac.kr), Dept. of Computer Science, Sangmyung University
• Received: 2019. 07. 17, Revised: 2020. 01. 03, Accepted: 2020. 01. 03.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

인터넷이 발달함에 따라 다양한 응용 프로그램과 미디어를 통해 하루에도 수십억 이상의 데이터가 생성된다. 이에 따라 방대한 데이터 중에서 정보를 요약하거나 추출하는 기술의 필요성이 강조되고 있다. 키워드 추출은 정보 검색, 문서 요약, 주제 탐지 등에 활용되는 기술로 텍스트 마이닝의 기반이 된다. 대용량 전자문서로부터 추출된 키워드들은 텍스트 분류, 텍스트 클러스트링, 기계 번역, 음성 인식 등의 텍스트 마이닝을 위한 주요한 속성으로 활용된다[1].

텍스트 랭크 알고리즘은 소셜 네트워크, 웹의 링크 구조 등의 분석에 사용된 그래프 기반의 랭킹 알고리즘과 자연어 처리를 결합한 기법이다. 텍스트를 그래프로 모델링 한 뒤 그래프 기반 랭킹 알고리즘을 적용한 후 핵심적인 정점을 추출해 내는 방법을 사용한다. 텍스트 랭크 알고리즘은 별도의 학습 데이터가 필요하지 않는 unsupervied learning 알고리즘이다[2]. 학습 데이터가 없거나 부족한 경우(실시간 정보 위주의 SNS와 뉴스 기사)에서 유용하게 사용된다. 또한 텍스트 랭크 알고리즘은 모듈화된 구성으로 이루어져 있다. 각 모듈을 수행하고자 하는 자연어 처리 작업의 특성에 맞게 수정하거나 대체할 수 있어 작업에 따라 적용할 수 있는 유연성이 장점이다[3]. 본 논문에서는 키워드 추출을 위한 텍스트 랭크 알고리즘의 성능을 분석한다. 다양한 분야에서 가장 보편적으로 활용되는 알고리즘 4가지의 활용 분야와 최신 기술을 설명하고 실험을 통한 분석을 실시한다.

가장 먼저, TF-IDF[4]가 있다. TF-IDF는 어떤 단어가 쿼리에서 사용하기에 더 효율적인지 결정하기 위해 개발되었다. TF-IDF는 문서 내에서 특정 단어의 빈도를 전체 문서의 단어 출현 빈도로 나눈 값이다. TF-IDF는 벡터 공간 모델로 비정형적인 텍스트 문서를 단순하고 정화된 모델로 표현함으로 기존 데이터 마이닝에서 사용되었던 알고리즘들을 수정 없이 적용할 수 있는 특징이 있다. 이러한 특징으로 현재까지도 많은 연구에서 활용되고 있다.

다음으로 SMART[5]가 있다. SMART는 하버드와 코넬 대학에서 30년이 넘는 기간 동안 개발된 검색 시스템으로 벡터 공간 및 정보 검색과 관련된 모델을 조사하기 위한 프레임 워크를 제공한다. SMART는 시간당 600Mbyte의 속도로 문서를 검색하며 간단한 벡터 검색 실행의 속도가 빠른 장점을 가지고 있다. 또한 다른 언어에 대한 배경 지식이 없어도 일부 외국어에 쉽게 적용된다.

INQUERY[6]는 베이지안 기법을 기반으로 하는 확률론적 정보 검색 시스템으로 사전 인덱싱 된 문서를 클러스트링하는 방법이다. 인터넷 상에서 많은 검색 사이트에 적용되었으며, 다국적 언어를 지원한다. 또한 INQUERY는 다양한 인덱싱 기술을 지원하고 정보 필터링, 데이터베이스와의 연동 등 다양한 분야와 결합되어 활용되고 있다.

마지막으로 CCA[7]가 있다. CCA 알고리즘은 유전자 프로그래밍을 사용하여 랭킹 알고리즘에 도입한 접근법으로 가중치 구성 요소의 조합을 기반으로 수집된 문서에서 파생된 기본 통계 정보만을 사용하는 다른 알고리즘보다 검색 기능을 향상 시켰다. 또한 CCA는 다른 GP기반의 알고리즘 FAN-GP보다 빠르게 수렴하는 장점을 가졌다.

본 논문에서는 위에서 언급한 4가지의 알고리즘의 성능을 실험을 통해 비교 분석한다. 실험은 뉴스와 트위터의 데이터를 수집하여 수집된 데이터 내에서 각 알고리즘을 통해 추출된 토픽을 분석하였다. 또한 미리 설정한 토픽을 기준으로 Rand statistic[8], Jaccard coefficient[9], FM index[10], Odds Ratio[11], Relative risk[12] F-measure[13] 여섯가지의 메트릭을 통해 각 알고리즘의 성능을 비교하였다. 실험 결과로 뉴스데이터는 5가지 메트릭 모두에서 CCA가 가장 좋은 성능을 보였다. 가장 낮은 성능을 보인 알고리즘은 Inquery로 Odds Ratio 에선 CCA는 5.120, Inquery는 0.419로 10배 이상의 차이를 보였다. 트위터 데이터의 경우에서는 CCA와 TF-IDF가 비슷한 성능을 보이며, SMART와 Inquery보다 좋은 성능을 보였다.

본 논문의 구성은 다음과 같다. 2장에서는 텍스트 랭크 알고리즘에 대해 설명한다. 3장에서는 각 알고리즘의 성능을 분석하기 위하여 실험을 진행한다. 4장에서는 실험의 결과를 확인한다. 5장은 마지막 결론 순으로 구성되어 있다.

II. Algorithms

1. TF-IDF

TF-IDF는 문서나 문장 내에서, 특정 단어가 해당 문서 내에서의 중요도를 통계적 수치로 나타낸 것이다. 문서 내에서 특정 단어의 빈도를 전체 문서의 단어 출현 빈도로 나눈 값을 수치로 표현한 것으로 다음과 같은 식으로 나타낸다.

$$TF-IDF = tf(t,d) \times idf(t,D) \quad (1)$$

식 (1)에서 t는 특정 단어, d는 전체 문서를 의미하며 $tf(t,d)$는 문서 d에서 특정 단어 t의 빈도 수를 전체 문서 길이로 나눈 값으로 단어 t가 문서 d에서의 빈도를 구한 값이다[14]. $idf(t,D)$의 값은 전체 문서의 수 D를 해당 단어를 포함한 문서의 수로 나눈 뒤 로그를 취한 값으

로 역문서의 빈도를 구한 값이다. TF-IDF는 단어가 특정 문서 내에서 빈도수가 높고 전체 문서 중 해당 단어를 포함된 문서가 적을수록 높아지며 이를 통해 모든 문서에서 자주 나타나는 단어를 추출할 수 있다. TF-IDF는 키워드 추출과 검색 결과의 순위, 문서 유사도 등에 활용되며 데이터 마이닝 분야에서 활용되고 있다[15]. 먼저 CLEF 2019 학회에서 제안된 트위터 프로파일링 작업의 성능을 향상 시키기 위해 TF-IDF가 사용된 연구가 있다[16]. 슬로바키아 공과대학 연구진은 랜덤 포레스트 분류기와 TF-IDF의 기능 추출방법을 기반으로 텍스트 정규화를 위한 방법을 제안했으며, 그 결과 정확성이 증가하였고, 처리 시간이 단축되었다. 다음으로 Guo, Jinghuan, et al는 사물인터넷의 일반적인 고용 해결 전략을 개선하기 위한 TF-IDF기반의 활동 특징 해결을 제안했다. 이를 통해 활동 인식 시스템의 성능을 획기적으로 향상시켰다[17].

2. SMART

SMART는 G. Salton의지도하에 개발된 실험 정보 검색 시스템이다. 스마트는 벡터 공간 모델을 기반으로 구문 분석, 단어 출현에 대한 통계적 분석을 통해 생성된 구문, 단어들을 벡터로 다음과 같이 표현한다.

$$d_i = (wi_1, wi_2, \cdots, wi_k) \tag{2}$$

식 (2)에서 d_i는 문서를 의미한다. wi_k는 문서 내에서 특정 단어 t_k의 가중치를 의미하며 문서 내에서 중요도를 나타낸다. 식 (2)와 같이 벡터가 형성되면, 검색 과정은 벡터 연산에 의하여 이루어진다. 벡터의 연산은 벡터들의 내적으로 계산되며, d_i의 값은 가중치 t_k에 의해 결정되기 때문에 SMART에서 가중치 부여 기법은 성능에 중요한 영향을 미친다[18]. 따라서 SMART에서는 다음과 같은 식을 활용하여 가중치를 계산한다.

$$wi_k = \frac{(\log(f_{ik})+1.0)\times\log(N/n_k)}{\sqrt{\sum_{j=1}^{t}[(\log(f_{ij})+1.0)\times\log(N/n_j)]^2}} \tag{3}$$

식(3)에서 tf_{ik}는 문서 내에서의 특정단어 t_k의 출현 빈도를 의미하며, N은 전체 문서의 수, 그리고 n_k는 특정 단어 t_k가 포함된 문서의 수를 나타낸다. 식 (3)은 특정 단어 출현 빈도와 역 문서 빈도를 곱한 값을 코사인 정규화 한 것이다[19]. 식 (3)에서 구한 벡터 값과 인덱싱 된 문서와 비교하여 유사성을 얻고 유사성을 기준으로 문서의 순위를 결정한다.

SMART는 언어에 대한 지식 없이도 외국어에 쉽게 적용될 수 있으며, 벡터 공간 및 정보 검색과 관련된 모델을 조사하기 위한 기본 프레임 워크를 제공한다. SMART는 Information Retrieval(IR) 시스템 내에서 효과적인 검색을 위한 쿼리 작업으로 처음 제안된 이후, 이를 개선하기 위한 많은 연구가 있었다. 관련 개념을 식별하기 위한 오류 또는 온톨리지의 적용 연구와[20], 웹 검색에서 쿼리 작업을 위해 현재까지 활용되고 있다[21].

3. INQUERY

INQUERY는 매사추세츠 대학에서 개발된 문서를 클러스터링하는 방법으로 이미 정의된 용어 사전을 이용하여 개념화한 후 Bayesian 네트워크를 사용하여 개념을 문서와 매칭 시키는 방법이다. INQUERY에서 각 문서에 대해 확률론적 값을 계산하는 식은 다음과 같다.

$$w_{t,d} = 0.4+0.6\times\frac{tf_{t,d}}{tf_{t,d}+0.5+1.5\frac{length_{(d)}}{avg\,len}}\times\frac{\log\frac{N+0.5}{n_t}}{\log N+1} \tag{4}$$

식 (4)에서 n_t는 특정 단어 t를 포함하는 문서의 수를 의미하고, N은 전체 문서의 수, $avglen$은 전체 문서의 평균 길이, $length_{(d)}$는 문서의 길이를 의미한다. tf는 문서 d에서 특정 단어 t의 빈도 수를 의미한다[22].

INQUERY의 주요 프로세스는 문서 인덱싱, 쿼리 처리, 쿼리 평가로 간단한 단어 기반 인덱싱, 품사 태그 지정에 기반한 인덱싱, 도메인 종속 기능에 의한 인덱싱을 비롯한 다양한 인덱싱 기술을 지원합니다. 이러한 인덱싱은 INQUERY의 토픽 탐지 및 추출 시스템을 제공한다. INQUERY는 인터넷 상에서 많은 검색 사이트에 적용되었으며, 다국적 언어를 지원한다. 최근에는 아랍어 텍스트 검색을 개선하기 위한 형태소 분석에도 사용되었으며[23], 또한 정보 필터링, 데이터베이스와의 연동 등 다양한 분야와 결합되어 활용되고 있다[24]. 대표적으로 Daou, Hoda.는 방대한 소셜 미디어 정보에서 관련된 정보를 추출하기 위해 INQUERY를 사용하였고, 그 결과 대중에게 강한 감정을 불러 일으키는 사건(총격 및 비행기 추락)과 같은 사건을 식별할 수 있게 되었다[25].

4. CCA

CCA 알고리즘은 Genetic Programming (GP)에 기반하여 수집된 문서 내에서 랭킹을 발견하는 새로운 방법으로 HM de Almeda et al.에 의해 처음 제안되었다. CCA는 순

위 함수에서 추출한 용어 가중치 요소(용어 빈도, 수집 빈도 정규화)의 조합으로 트레이닝과 확인을 반복하는 단계로 구성된 프로세스이다[26]. 알고리즘은 다음과 같다.

$$
\begin{aligned}
CCA = ((99.09+t_{09}) + (((t_{06}\times t_{08})\times \\
(t_{05}\times((((t_{06}\times t_{08})+(t_{07}\times t_{08}))\times \\
(t_{10}\times t_{01})))) \\
+((t_{06}\times t_{08})\times(t_{05}\times(((t_{02}\times t_{04})+(t_{07}+t_{08})) \\
\times(t_{10}\times t_{01}))))))+((t_{10}\times t_{01}) \\
+((t_{06}\times t_{08})\times(t_{05}\times(((t_{07}\div t_{03})+(t_{07}+t_{08})) \\
\times(t_{10}\times t_{01})))))
\end{aligned} \tag{5}
$$

식 (5)에서 t_i가 나타내는 수식은 표1에 나와 있다. t_{01}은 문서에서 특정 단어의 빈도 수를 의미하며, t_{02}는 tf의 자연대수, t_{03}는 최대 tf로 정규화된 tf 계수를 의미한다. t_{04}는 SMART 가중치 수식, t_{05}는 Okapi BM25 수식을 의미하며 이때 $k_1=1000$이다. t_{06}는 역 문서 빈도(idf)의 대안이며, t_{07}는 Robertson Sparck Jones의 변형이다. t_{08}는 확률적 역 수집 빈도와 t_{09}는 INQUERY 수식을 나타낸다. t_{10}은 Cosine 정규화이며 $w_{i,j}^2 = t_{01}\times t_{06}$이다. 마지막으로 t_{11}은 문서 길이의 정규화이다 [7]. CCA 알고리즘은 기존의 랭킹 알고리즘 정확도가 개선되었으며, 유전 프로그래밍에서 발생하는 문제점인 초과학습을 줄였다. 그 결과로 다른 GP 기반 방식의 알고리즘보다 속도를 향상시켰다.

Baeza-Yates, Ricardo, et al는 CCA 알고리즘을 활용하여 웹 문서 순위를 지정하는 효과적인 알고리즘 CombGenRank 알고리즘을 제안하였고 클라우트 컴퓨팅을 기반으로 하는 검색 서비스에서 효과를 보였다[27]. 또한 CCA는 정보 검색에서 용어 가중치 접근법의 문제를 해결하기 위한 토대로 사용되고 있다[28].

III. Experiment

본 장에서는 2장에서 언급된 알고리즘들의 성능을 비교하기위하여 실험을 진행한다. 먼저 실험에 사용된 데이터는 네이버 뉴스와 트위터에서 데이터를 수집하였다. 2018년 8월 1일부터 15일까지 대한민국에서 발생된 '폭염'과 관련된 데이터를 수집한 후에 각 알고리즘들을 사용하여 폭염의 랭킹과 이와 관련된 데이터 추출 비율을 비교하였다. 표2는 전체 데이터, 단어의 수, 중복이 제거된 단어의 수, 바이트로 데이터 셋을 보여준다. 우리는 각 알고리즘의 성능 평가를 위해 6가지의 평가 척도(Rand statistic, Jaccard coefficient, FM index, Odds Ratio, Relative risk, F-measure)를 사용했다.

Table 1. CCA Algorithm Formula

ID	Equation
t_{01}	tf
t_{02}	$1+\log(tf)$
t_{03}	$0.5+\frac{0.5+tf}{\max tf}$
t_{04}	$\frac{1+\log(tf)}{1+\log(avg\,tf)}$
t_{05}	$\frac{(k_1+1)\times tf}{(k_1((1-b)+b\times dl/avg\,dl)+dl)+tf}$
t_{06}	$\log(\frac{N}{df}+1)$
t_{07}	$\log(\frac{N-df+0.5}{0.5})$
t_{08}	$\log(\frac{N-df}{df})$
t_{09}	$\frac{\log\frac{N+0.5}{df}}{\log N+1}$
t_{10}	$\frac{1}{\sqrt{\sum_{i=0}^{2} w_{i,j}^2}}$
t_{11}	dl

6가지의 메트릭은 식(6)-(11)과 같이 정의된다. 여기서 TP는 발생한 이벤트와 관련된 뉴스 및 트위터 데이터의 수이고, TN은 발생하지 않은 이벤트와 관련이 없는 뉴스 및 트위터 데이터의 수이고, FP는 발생하지 않은 이벤트와 관련된 뉴스 및 트위터 데이터의 수이다. FN은 발생한 이벤트와 관련이 없는 뉴스 및 트위터 데이터의 수를 의미한다. TP와 TN은 'positive topic', FP와 FN은 'negative topic'이다.

Table 2. Statistics of the data set
(Average value from 2018.08.01. - 2018.08.15.)

Data Set	News	Tweets
Total Data	1,982	5,084
Terms	55,229	51,107
Bytes	732,826	1,062,958
Unique Terms	9,526	11,689

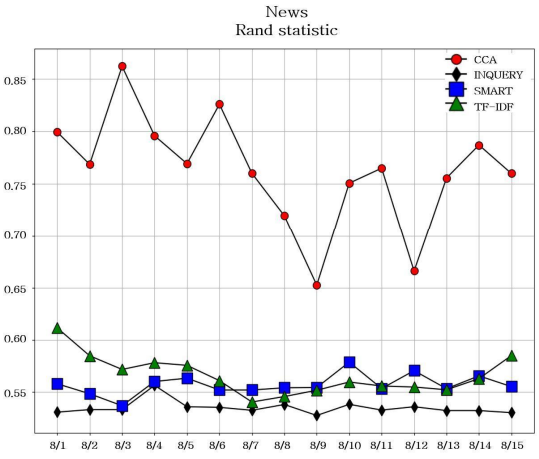

Fig. 1. Rand Statistic Result Graph - News

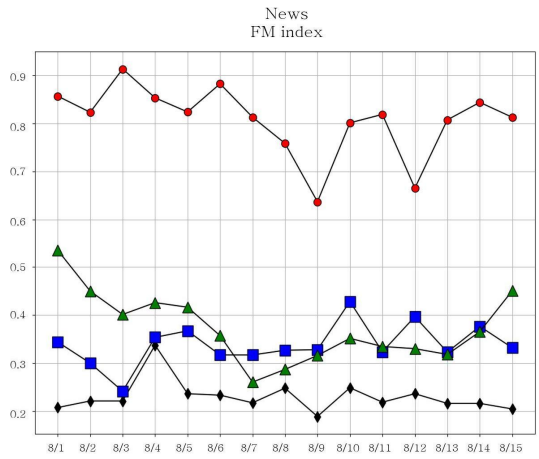

Fig. 2. FM index Result Graph - News

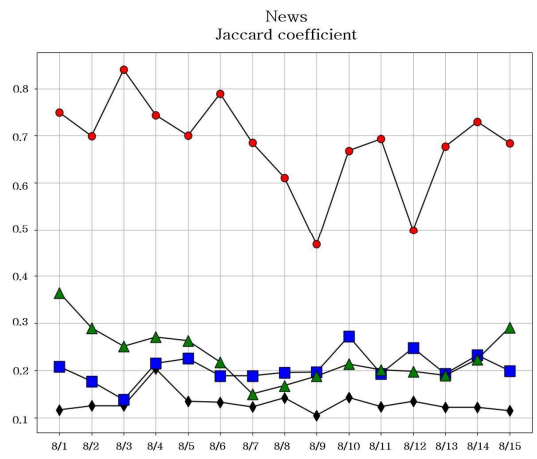

Fig. 3. Jaccard coefficient Result Graph - News

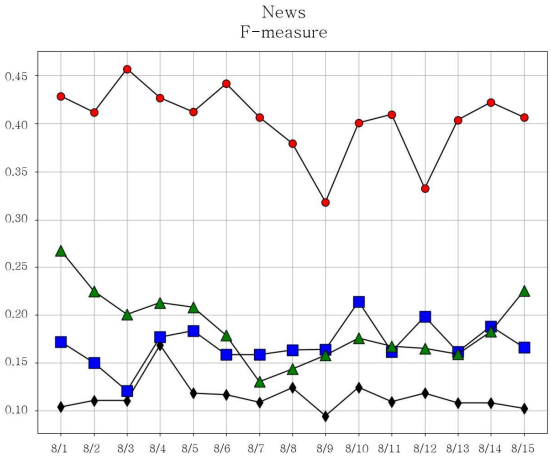

Fig. 4. F-measure Result Graph - News

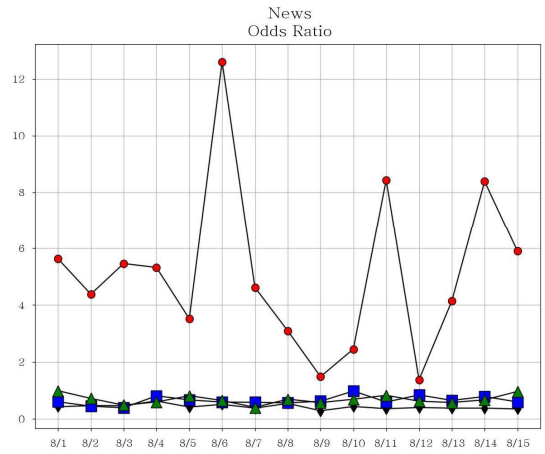

Fig. 5. Odds Ratio Result Graph - News

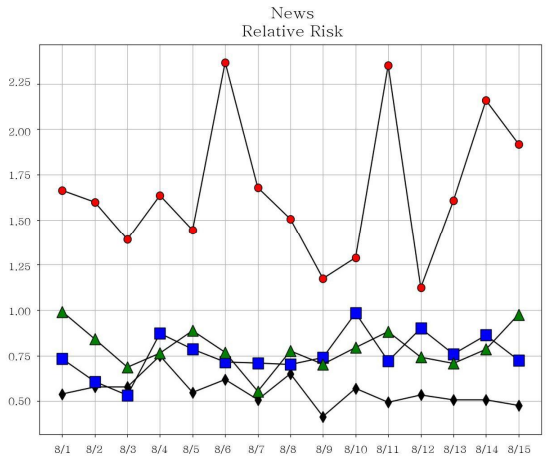

Fig. 6. Relative Risk Result Graph - News

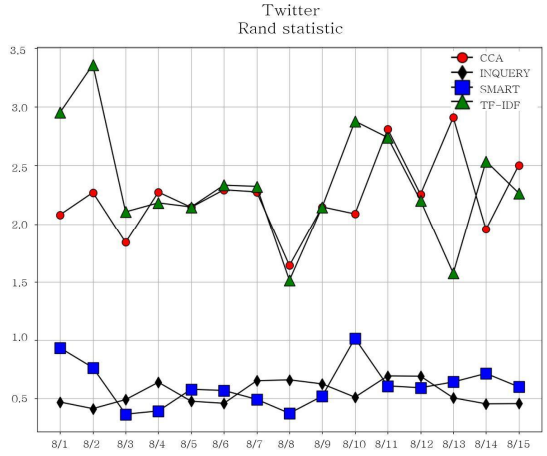

Fig. 7. Rand Statistic Result Graph - Twitter

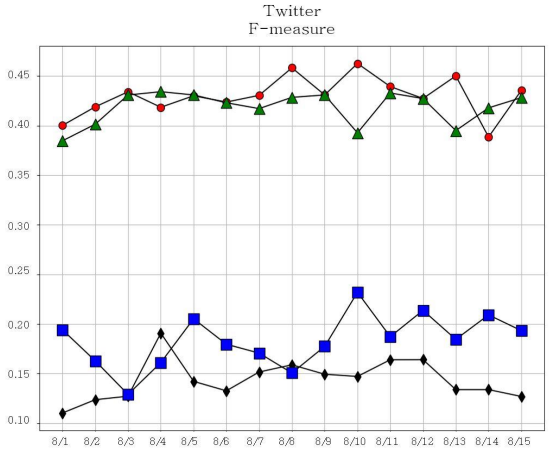

Fig. 10. F-measure Result Graph - Twitter

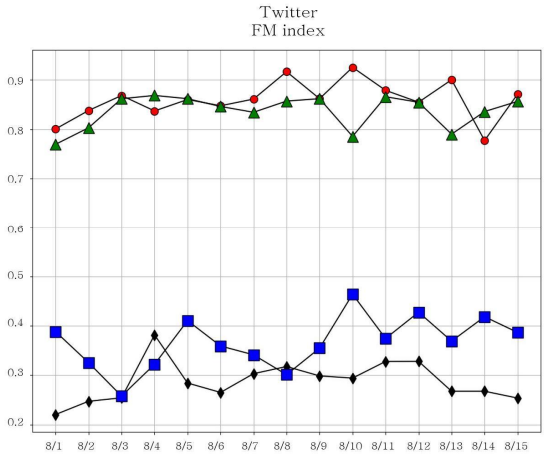

Fig. 8. FM index Result Graph - Twitter

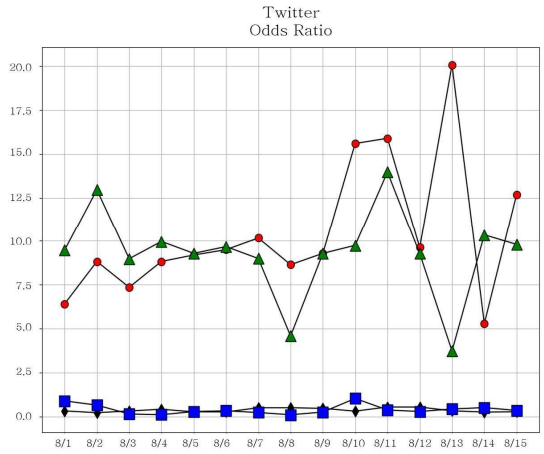

Fig. 11. Odds Ratio Result Graph - Twitter

Fig. 9. Jaccard coefficient Result Graph - Twitter

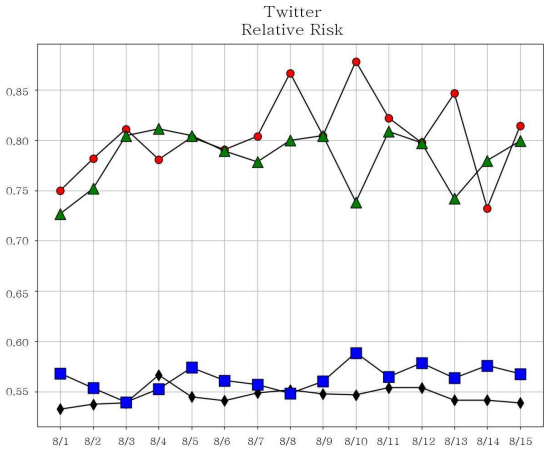

Fig. 12. Relative Risk Result Graph - Twitter

Table 3. Ranking of specific words ('폭염') by period

	TF-IDF		SMART		INQUERY		CCA	
	news	twitter	news	twitter	news	twitter	news	twitter
20180801	3	118	4	125	7	234	8	101
20180802	1	76	3	79	6	143	8	67
20180803	1	138	3	142	6	290	8	108
20180804	1	312	2	327	6	601	8	277
20180805	1	145	2	160	5	368	9	107
20180806	2	117	3	126	6	291	8	85
20180807	3	349	3	368	6	646	9	321
20180808	4	262	4	269	10	448	5	268
20180809	3	282	5	310	10	733	4	209
20180810	3	367	6	394	10	734	4	314
20180811	5	397	8	397	11	651	9	384
20180812	10	755	9	793	13	1426	9	692
20180813	4	399	6	417	11	785	4	355
20180814	11	289	9	315	17	570	8	262
20180815	14	558	13	593	18	1176	12	426

Table 4. Average value of the algorithm News data

NEWS	TF-IDF	SMART	INQUERY	CCA
Rand statistic	0.566	0.557	0.535	0.762
Jaccard coefficient	0.232	0.205	0.130	0.682
FM index	0.373	0.807	0.230	0.807
Odds ratio	0.677	0.640	0.419	5.120
Relative risk	0.790	0.756	0.551	1.66
F-measure	0.186	0.169	0.115	0.403

Table 5. Average value of the algorithm Twitter data

TWITTER	TF-IDF	SMART	INQUERY	CCA
Rand statistic	0.782	0.563	0.546	0.805
Jaccard coefficient	0.720	0.225	0.168	0.756
FM index	0.836	0.366	0.287	0.859
Odds ratio	9.332	0.364	0.395	10.485
Relative risk	2.351	0.608	0.554	2.232
F-measure	0.418	0.183	0.143	0.429

$$Rand\,statistic = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Jaccard\,coefficient = \frac{TP}{TP+FP+FN} \quad (7)$$

$$FM\ Index = \frac{TP}{\sqrt{TP+FP}+\sqrt{TP+FN}} \quad (8)$$

$$Odds\ Ratio = \frac{TP\times TN}{FP\times FN} \quad (9)$$

$$Relative\ Risk = \frac{TP/(TP+FP)}{FN/(FN+TN)} \quad (10)$$

$$Dice\ \ Index = \frac{2\times TP}{2\times TP+FP+FN} \quad (11)$$

Rand statistic은 positive topic과 negtive topic과 같은 두 데이터 집합의 유사성을 측정하는데 사용된다. Jaccard coefficient는 데이터 집합의 유사성과 다양성을 비교하는 측정할 수 있으며, 이를 통해 정확도를 분석한다. FM index 두 데이터 집합 간의 유사성을 결정하는 측정 기술이며, Odds Ratio는 positive topic과 negtive topic 간의 연관성을 나타내는 척도이다. Relative risk는 두 데이터 집합 간의 수치를 바탕으로 직관적인 차이를 분석한다. F-measure은 정보 검색분야에서 검색, 문서분류 및 쿼리 분류 성능을 측정하는데 사용된다.

IV. Result

표3은 알고리즘에 대한 특정 단어 '폭염'의 랭킹 순위로 각 알고리즘의 특정 단어의 추출 성능을 보여준다. 2018년

8월 1일부터 15일까지의 평균 순위는 TF-IDF가 4위로 가장 높은 순위였으며, INQUERY가 9위로 가장 낮은 순위로 4가지 알고리즘 모두 뉴스 데이터에서 폭염은 높은 순위를 차지했다. 이는 모든 알고리즘이 뉴스 데이터에서 특정 단어 관련 정보를 추출 성능이 높다는 것을 알 수 있다. 하지만 트위터 데이터에서는 알고리즘의 차이가 두드려졌다. 트위터의 경우, CCA 알고리즘이 265로 가장 높은 순위를 차지했으며, INQUERY가 606위로 2배 이상의 차이를 보였다. 이는 뉴스 데이터보다 트위터 데이터에서 알고리즘들의 성능차이가 나타나는 것을 확인할 수 있으며, 뉴스 데이터가 트위터보다 데이터 집합의 질이 더 좋은 것을 알 수 있다.

다음으로 그림 1-6와 표4는 뉴스 데이터 기반의 2018년 8월 1일부터 15일까지의 Rand statistic, Jaccard coefficient, FM index, Odds Ratio, Relative Risk, F-measure에 대한 알고리즘의 결과를 보여준다. 그래프와 표에서 확인 할 수 있듯이 뉴스 데이터의 경우 CCA가 다른 알고리즘들 보다 훨씬 우수한 성능을 보여준다. Odds Ratio의 경우 CCA 알고리즘은 가장 높은 5.120 이였으며, INQUERY는 0.419로 가장 낮은 성능을 보이며 10배 이상의 차이를 보였다. CCA를 제외한 다른 3가지의 알고리즘은 비슷한 수준을 보였고, 그 중 TF-IDF가 가장 좋은 성능을 보였다.

마지막으로 그림 7-12과 표5은 트위터 데이터 기반의 2018년 8월 1일부터 15일까지의 Rand statistic, Jaccard coefficient, FM index, Odds Ratio, Relative Risk, F-measure에 대한 알고리즘의 결과를 보여준다. 트위터의 경우 CCA가 평균 Rand statistic은 0.805, Jaccard coefficient은 0.756, FM index은 0.859, Odds Ratio은 10.485, Relative Risk은 2.232, F-measure은 0.429 였으며, TF-IDF는 평균 Rand statistic은 0.782, Jaccard coefficient은 0.720, FM index은 0.836, Odds Ratio은 9.332, Relative Risk은 2.351 F-measure은 0.418로, SMART, INQUERY 알고리즘과 비교하여 우수한 성능을 보였다.

결과적으로 CCA는 뉴스와 트위터 데이터 모두에서 가장 우수한 성능을 보였으며, TF-IDF는 뉴스보다 트위터에서 더 좋은 성능을 보인다. 다음으로 SMART INQUERY 알고리즘의 경우, 뉴스 데이터에서 특정 단어추출 성능은 높지만 뉴스와 트위터 데이터에서 추출한 단어에 따른 정확성이 높지 않다는 것을 확인 할 수 있었다.

V. Conclusions

텍스트 랭킹 알고리즘의 중요성은 지속적으로 증가하고 있으며, 다양한 연구들이 진행되었다. 하지만 기존의 텍스트 랭킹 알고리즘들의 비교가 부족하였고, 본 논문에서는 실험을 통하여 대표적인 텍스트 랭크 알고리즘의 성능을 분석했다. 뉴스 데이터의 경우 특정 단어를 추출하는 성능은 4가지 알고리즘 모두 우수한 성능을 보였으며, 트위터의 경우 CCA가 가장 좋은 성능을 보였으며, TF-IDF, SMART, INQUERY 순으로 추출 능력을 보였다. 또한 정확성을 비교하기 위한 실험을 통해 뉴스 데이터와 트위터 데이터 모두에서 CCA가 가장 높은 정확성을 보였으며, TF-IDF는 뉴스보다 트위터에서 더 좋은 성능을 보였다. 대용량의 문서에서 키워드 추출 기술의 필요성이 꾸준히 증가함에 따라 텍스트 랭크 알고리즘 또한 지속적으로 연구되고 있으며, 더욱 많은 분야에서 활용될 것이며 연구의 가치가 크다.

REFERENCES

[1] B. Lott, “Survey of keyword extraction techniques,” UNM Education, vol. 50, pp. 1-11, 2012.

[2] Y. K. Meena, A. Jain, and D. Gopalani, “Survey on graph and cluster based approaches in multi-document text summarization,” in International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), 2014, pp. 1-5.

[3] S. S. Sonawane and P. A. Kulkarni, “Graph based representation and analysis of text document: A survey of techniques,” International Journal of Computer Applications, vol. 96, no. 19, 2014.

[4] Ramos, “Using tf-idf to determine word relevance in document queries,” in Proceedings of the first instructional conference on machine learning, 2003, vol. 242, pp. 133-142.

[5] C. Buckley, G. Salton, J. Allan, and A. Singhal, “Automatic query expansion using SMART: TREC 3,” NIST special publication sp, pp. 69-69, 1995.

[6] J. P. Callan, W. B. Croft, and S. M. Harding, “The INQUERY retrieval system,” in Database and expert systems applications, 1992, pp. 78-83.

[7] H. M. de Almeida, M. A. Gonçalves, M. Cristo, and P. Calado, “A combined component approach for finding collection-adapted ranking functions based on genetic programming,” in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 399-406.

[8] D. G. Fisher and P. Hoffman, “The adjusted Rand statistic: A

SAS macro," Psychometrika, vol. 53, no. 3, pp. 417-423, 1988.
[9] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of Jaccard coefficient for keywords similarity," in Proceedings of the international multiconference of engineers and computer scientists, 2013, vol. 1, pp. 380-38
[10] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," Journal of intelligent information systems, vol. 17, no. 2-3, pp. 107-145, 2001.
[11] C. O. Schmidt and T. Kohlmann, "When to use the odds ratio or the relative risk?," International journal of public health, vol. 53, no. 3, pp. 165-167, 2008.
[12] J. Zhang and F. Y. Kai, "What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes," Jama, vol. 280, no. 19, pp. 1690-1691, 1998.
[13] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).
[14] Z. Yun-tao, G. Ling, and W. Yong-cheng, "An improved TF-IDF approach for text classification," Journal of Zhejiang University-Science A, vol. 6, no. 1, pp. 49-55, 2005.
[15] T. Roelleke and J. Wang, "TF-IDF uncovered: a study of theories and probabilities," in Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp. 435-442.
[16] Guo, Jinghuan, et al. "Activity feature solving based on TF-IDF for activity recognition in smart homes." Complexity 2019 (2019).
[17] Petrik, Juraj, and Daniela Chuda. "Twitter Feeds Profiling With TF-IDF." CLEF. 2019.
[18] Kyi Ho Lee, Joon Ho Lee, Kyu Chul Lee., "Improving Retrieval Effectiveness with Multiple Query Combination," JOURNAL OF THE KOREAN SOCIETY FOR LIBRARY AND INFORMATION SCIENCE 31(3), 1997.9, 135-146(12 pages)
[19] C. Buckley, A. Singhal, M. Mitra, and G. Salton, "New retrieval approaches using SMART: TREC 4," in Proceedings of the Fourth Text REtrieval Conference (TREC-4), 1995, pp. 25-48.
[20] Macdonald, Craig, Nicola Tonellotto, and Iadh Ounis. "Efficient & effective selective query rewriting with efficiency predictions." Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2017.
[21] Raza, Muhammad Ahsan, et al. "A Taxonomy and Survey of Semantic Approaches for Query Expansion." IEEE Access 7 (2019): 17823-17833.
[22] J. P. Callan, W. B. Croft, and J. Broglio, "TREC and TIPSTER experiments with INQUERY," Information Processing & Management, vol. 31, no. 3, pp. 327-343, 1995.
[23] Nwesri, Abdusalam F. Ahmad, and Hasan AH Alyagoubi. "Applying Arabic Stemming Using Query Expansion." 2015 26th International Workshop on Database and Expert Systems Applications (DEXA). IEEE, 2015.
[24] J. Allan, L. Ballesteros, J. P. Callan, W. B. Croft, and Z. Lu, "Recent experiments with INQUERY," in Proceedings of the 4th Text Retrieval Conference, 1995, pp. 49-64.
[25] Daou, Hoda. "Detection of Sentiment Provoking Events in Social Media." Proceedings of the 52nd Hawaii International Conference on System Sciences. 2019.
[26] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," IEEE Transactions on Speech and Audio Processing, vol. 11, no. 3, pp. 204-215, 2003.
[27] Baeza-Yates, Ricardo, et al. "An effective and efficient algorithm for ranking web documents via genetic programming." Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. ACM, 2019.
[28] Fernández, Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. "Learning to Weight for Text Classification." IEEE Transactions on Knowledge and Data Engineering (2018).

Authors

Myung-Hwi Kim received the B.S. degree in computer science from Sangmyung University, Seoul, South Korea, in 2019, where he is currently pursuing the M.S. degree with the Department of Computer Science. His research interests include machine learning and computer networks.

Beakcheol Jang received the B.S. degree from Yonsei University in 2001, the M.S. degree from Korea Advanced Institute of Science and Technology in 2002, and the Ph.D. degree from North Carolina State University in 2009, all in Computer Science. Dr. Jang joined the faculty member of the department of Media software at sangmyung University, Seoul, Korea, in 2012. He is currently an assistant professor in the Department of Media Software, Sangmyung Univerisy. He is interested in wireless networking with an emphasis on ad hoc networking, wireless local area networks, and mobile network technologies.