[UCI]I804:24011-200000266010

<별첨1> 표지

2017년 2월
석사학위 논문

문서 내 단어 분포차를 고려한 개선된 범주 분류 방법

조선대학교 대학원

컴퓨터공학과

이 재 욱

2017년2월
석사학위논문
문서 내 단어 분포차를 고려한 개선된 범주 분류 방법
이재욱

2017년 2월

석사학위 논문

문서 내 단어 분포차를 고려한 개선된 범주 분류 방법

조선대학교 대학원

컴퓨터공학과

이 재 욱

문서 내 단어 분포차를 고려한 개선된 범주 분류 방법

An improved category classification method considering word distribution difference in the document

2017년 2월 24일

조선대학교 대학원

컴퓨터공학과

이 재 욱

문서 내 단어 분포차를 고려한 개선된 범주 분류 방법

지도교수 김 판 구

이 논문을 공학석사학위 신청 논문으로 제출함

2016년 10월

조선대학교 대학원

컴퓨터공학과

이 재 욱

이재욱의 석사학위논문을 인준함

위원장 조선대학교 교수 정 일 용 (인)

위 원 조선대학교 교수 김 판 구 (인)

위 원 조선대학교 교수 양 희 덕 (인)

2016년 11월

조선대학교 대학원

목 차

Ⅰ. 서론 ····· 1
1. 연구 배경 및 목적 ····· 1
2. 논문의 구성 ····· 3

Ⅱ. 관련 연구 ····· 4
1. 키워드 추출 방법 ····· 4
1) 이진 가중치 ····· 5
2) 단어 빈도 ····· 5
3) TF-IDF ····· 6
4) TF-ICF ····· 7
5) TF-ISF ····· 8
2. 문서 유사도 측정 ····· 9
1) 코사인 유사도(Cosine Similarity) ····· 9
2) 문서 빈도(Document Frequency) ····· 10
3) 카이제곱 통계량(X^2 Statistics) ····· 11
4) 상호 정보량(Mutual Information) ····· 12
5) 정보 획득량(Information Gain) ····· 13

Ⅲ. 문서 내 단어 분포를 고려한 키워드 가중치 측정 ····· 14
1. 전처리 ····· 15
1) 토크나이징(Tokenizing) ····· 15
2) 불용어(Stop word) 제거 ····· 15
3) 스태밍(Stemming) ····· 16
4) 품사 태깅(POS Tagging) ····· 17
5) 명사 추출 ····· 19

2. 위키피디아를 이용한 단어 확장 방법 ····· 20
1) 위키피디아를 이용한 정보 추출 ····· 20
2) Redirect 기능을 이용한 대표 단어 추출 ····· 21

3. 단어의 분포를 이용한 가중치 측정 방법 ····· 23
1) 단어 분포 편차가 큰 경우 ····· 24
2) 단어 분포 편차가 작은 경우 ····· 25
3) 모든 문서를 고려한 편차 측정 방법 ····· 27

Ⅳ. 실험 및 평가 ····· 30
1. 실험 데이터 ····· 30
2. 카테고리 별 사전데이터 구축 ····· 31
3. 표준편차를 적용한 키워드 가중치 측정 ····· 32
4. 비교 실험 ····· 35

Ⅴ. 결론 및 향후연구 ····· 38

【참고문헌】 ····· 39

표 목 차

표 1. 같은 의미이지만 형태가 다른 단어 예시 ····· 2
표 2. 불용어 목록 예시 ····· 16
표 3. 스태밍 작업 예시 ····· 16
표 4. 영어 단어의 품사기호 ····· 17
표 5. 품사 태깅 알고리즘 ····· 18
표 6. 입력 문장을 대상으로 품사 태깅한 예시 ····· 19
표 7. 위키피디아의 Redirect 단어 ····· 21
표 8. 단어 car의 표준 편차 값 ····· 24
표 9. 단어 bus의 표준 편차 값 ····· 25
표 10. 단어의 편차 값 적용 방법 ····· 28
표 11. Entertainment 카테고리 단어의 표준 편차 값을 적용한 가중치 예 ····· 34
표 12. 실험 문서에서 추출된 명사 예시 ····· 35
표 13. TF-IDF를 이용한 실험 문서의 카테고리 분류 예시 ····· 36
표 14. 본 논문 방법을 이용한 실험 문서의 카테고리 분류 예시 ····· 36
표 15. TF-IDF의 문서 분류 결과 ····· 37
표 16. 본 논문의 문서 분류 결과 ····· 37

그 림 목 차

그림 1. 단어 A, B의 가중치 측정 예시 ····· 3
그림 2. 문서의 코사인 유사도 예시 ····· 10
그림 3. 전체 흐름도 ····· 14
그림 4. 위키피디아 문서 정보 예시(Businessperson) ····· 20
그림 5. Businessperson 단어의 redirect 정보 ····· 22
그림 6. 단어의 분포 편차가 큰 경우 ····· 24
그림 7. 단어의 분포 편차가 적은 경우 ····· 25
그림 8. 분포가 큰 단어의 표준편차 가중치 ····· 26
그림 9. 분포가 작은 단어의 표준편차 가중치 ····· 26
그림 10. 단어가 출현하지 않는 문서의 표준편차 측정 방법 ····· 27
그림 11. 단어 순위의 분포 차를 고려한 가중치 적용 알고리즘 ····· 29
그림 12. 한국 타임즈 신문에서 추출한 문서 집단 예 ····· 30
그림 13. 위키피디아를 이용한 사전데이터 구축 예 ····· 31
그림 14. 카테고리 별 단어 가중치 값 예시 ····· 32
그림 15. 카테고리 별 단어 순위화 ····· 33

ABSTRACT

An improved category classification method considering word distribution difference in the document

JaeUk Lee
Advisor : Prof. Pankoo Kim, Ph.D.
Department of Computer Engineering
Graduate School of Chosun University

Recently, the population of users of Social Network Service has been increasing due to the development of smart devices that is where big-data be are accumulated. Pattern recognition, machine learning, nature language processing can process and analyze the big-data. Nature language processing is one of the methods which deals with human language using computer. This area focuses on a human availability, and many other applications such as keyword extraction, information retrieval, document summarization, document classification and so on about a human availability.

This paper proposed the method of extracting keyword that special feature words assign high score and without discernment words assign low score in document. First, we extract one of the representative word that same meaning and different shape words such as neologism, abbreviation and synonym word using 'redirect' option of Wikipedia. And then, we calculate the deviation to consider of all documents. Principal keyword have high weight value by as signing additional values to words above threshold value.

As a result, the precision rate has been increased up to 1.15% than the TF-IDF method.

Ⅰ. 서론

1. 연구 배경 및 목적

스마트 기기의 발달로 인해 인터넷 사용자들은 때와 장소를 가리지 않고 소셜 네트워크 서비스(Social Network Service)를 이용할 수 있게 되었다. 이로 인해 인터넷 안에서 거대한 정보와 데이터가 생성되었으며, 이를 빅 데이터(big data)라고 부른다. 빅 데이터는 데이터의 수집, 관리, 처리를 위한 소프트웨어가 수용할 수 있는 크기의 한계를 넘어서는 데이터를 말하며[1], 크기는 수십 테라바이트에서 수 페타바이트[1] 까지 이른다. 이처럼 방대하고 다양한 규모의 데이터는 미래의 중요한 데이터 자원으로 사용될 수 있기때문에 주목받고 있으며, 빅 데이터를 분석하고 효율적으로 관리하기 위해 데이터센터[2]라는 시설을 만들어서 관리하고 있다.

빅 데이터를 분석하는 연구 방법으로는 패턴 인식[3], 기계 학습[4], 자연어 처리 등이 있다[3][4][5]. 이 중 자연어 처리는 컴퓨터를 이용해 인간의 언어를 이해 및 분석하는 인공 지능 기술이다[6][7]. 자연어 처리는 주로 텍스트 문서를 분석하여 대표 키워드를 추출하고, 추출된 키워드로 정보검색, 자동번역, 질의응답, 문서분류 등을 한다. 이러한 빅 데이터 분석 기술을 이용한 대표적인 사례로는 버락 오바마가 2008년 미국 대통령 선거에서 유권자 데이터베이스를 분석한 정보를 수집한 후, 유권자 맞춤형 선거 전략으로 효과적인 선거를 수행한 적이 있고, 구글

---

1) 테라바이트(Terabyte, TB)는 10^{12}, 페타바이트(Petabyte, PB)는 10^{15}를 의미하는 SI(국제단위계) 접두어인 테라, 페타와 컴퓨터의 데이터를 표시하는 단위인 바이트가 합쳐진 자료량을 나타내는 단위이다. (1 TB = 10^{12} bytes, 1 PB = 10^{15} bytes)

2) 데이터센터는 빅데이터 분석 및 사물인터넷 구현 등 정보가 발생하는 모든 데이터를 저장하고, 저장된 데이터를 분석하는 곳을 말한다[2].

3) 패턴 인식(Pattern Recognition)은 공학적인 접근을 이용한 방법이며, 인공지능에서 구현 시, 대상을 센싱하여 인식하는 문제에 대해 주로 연구한다.

4) 기계 학습(Machine Learning)은 정보를 알고리즘으로 선행 학습하고, 선행 학습된 정보를 바탕으로 다른 정보를 효율적으로 적용하기 위한 분야이다.

은 자체적으로 축적한 데이터를 바탕으로 독감 트렌드를 분석하였고, 분석한 결과가 실제 독감 추세와 일치하여 빅 데이터 분야에서 큰 주목을 받은 연구 사례가 있었다[8].

방대한 데이터를 기반으로 서비스를 제공하는 SNS, 구글, 네이버와 같은 사이트는 일반 문서, 이미지 문서, 동영상 문서 등으로 카테고리를 나누어 정보를 제공하며, 사용자들은 보다 쉽게 정보를 검색하여 결과를 얻을 수 있다. 하지만 문서 안의 키워드가 불분명하여 사용자가 검색하는 키워드와 일치하지 않을 때에는 원하는 정보를 찾기 힘든 실정이다. 예를 들어 표 1과 같이 단어 옥수수의 방언인 옥고량, 옥시기 등의 생소한 단어로 정보를 검색할 경우에 모두 옥수수와 같은 동의어이지만, 문서 안에서 일치하는 단어의 형태가 달라 많은 정보를 찾지 못한다. 또한, 검색하려는 문서의 키워드가 내용이 전혀 다른 문서들 안에서 많이 쓰이게 된다면 불필요한 정보 검색 결과를 얻게 되는 문제가 있다.

표 1. 같은 의미이지만 형태가 다른 단어 예시

	예시
동의어[5]	옥수수, 강냉이, 찰옥수수, 강내미, 옥고량, 옥촉서, 진주미, 옥출, 직당, 포미, 옥미, 옥시기

본 논문에서는 각 문서를 대표하는 키워드가 불분명하여 정확한 키워드 매칭하지 못하고 TF-IDF에서 단어 빈도를 고려하지 못하는 문제점을 개선하는 방법을 제안한다. 먼저 온라인 백과사전인 위키피디아(Wikipedia) 정보를 이용해 각 문서 안의 신조어, 줄임말, 동의어 등의 단어에 대해 대표 단어를 추출하여 각 문서에 추가하고, 각 문서의 단어 정보를 바탕으로 단어 분포를 측정해 분포도가 큰 값 중에 문서를 대표 할 수 있는 평균 이상의 상위권 단어에 추가 가중치를 부여한다. 본 논문의 방법으로 적용된 단어 가중치의 효율성을 평가하기 위해 문서 내 단어 가중치 측정에 일반적으로 많이 쓰이는 방법인 TF-IDF(Term Frequency-Inverse Document Frequency)의 방법과 비교 평가한다.

5) 온라인 백과사전인 한글 위키피디아(wikipedia)에서 정의하고 있는 옥수수의 동의어.

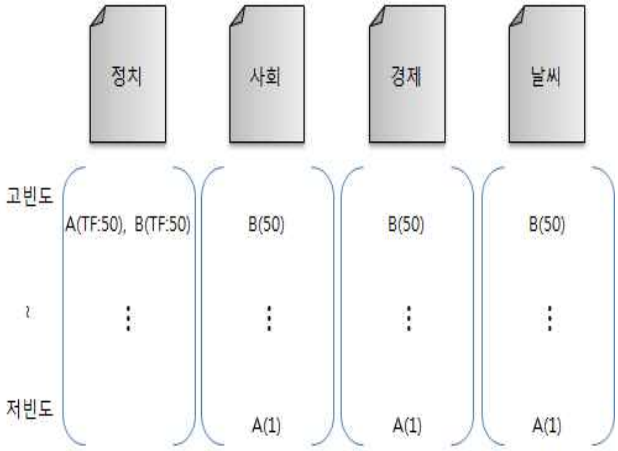


그림 2. 단어 A, B의 가중치 측정 예시

단어 B는 모든 카테고리에 균등하게 출현하여 특정 카테고리를 나타내기 힘들지만, 단어 A는 정치에만 높게 출현하여 정치에 특화된 단어이다. 하지만 TF-IDF 방법(2장 관련연구에서 자세히 다룸)을 이용해 정치 카테고리의 단어 A, B의 값을 측정하면 같은 값이 나온다. 왜냐하면 다른 카테고리에서 단어 A, B의 빈도차이를 고려하지 않았기 때문이다. 본 논문에서는 다른 카테고리의 단어 빈도차를 고려해 단어 A에 추가 가중치를 부여하는 개선된 방법을 제시한다.

2. 논문의 구성

본 장에 이어 관련 연구를 다룰 2장에서는 문서에서 키워드 추출하는 방법과 본 논문의 비교 실험에 쓰일 문서 분류 방법에 대한 관련 연구 및 현재 동향 대해 살펴본다. 3장인 본론에서는 기존의 단순 키워드 매칭으로는 찾을 수 없는 불분명한 키워드들에 대하여 위키피디아의 동의어 정보를 이용해 각 문서의 대표 단어를 추출하여 추가하고, 각 문서 내 단어 분포를 측정해 분포도가 평균 이상인 단어에 대해 추가 가중치를 부여하는 방법에 대해 상세히 기술한다. 4장인 실험 및 평가 부분에서는 3장에서 기술한 방법으로 각 문서의 단어 가중치를 측정하고 문서의 단어 가중치를 측정하는 방법인 TF-IDF 방법과 비교 실험한다. 마지막으로 5장에서는 본 연구의 전체적인 결론 및 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

가중치를 적용해 키워드를 추출하는 방법은 문서의 주제를 설명하기 가장 좋은 조건의 식별 방법이다[9]. 주로 텍스트 마이닝[6] 분야에서 키워드 추출을 하며 키워드 추출은 전처리[7] 작업 이후에 수행되는 연구의 기본이 되는 초기 단계 작업이다. 이 때문에 키워드 추출 결과에 따라 다음 단계에 지속적인 영향을 미친다.

본 장에서는 가중치를 적용해 키워드를 추출하는 방법 중 대표적으로 많이 쓰이는 키워드 추출 방법 및 수식에 대해 자세히 살펴본다. 또한, 국내외 관련 연구들을 정리하고, 특히 기존의 선행연구 중 TF-IDF를 이용해 문서 내 단어의 가중치를 적용한 연구를 분석하고 취약점을 도출해 이를 개선할 수 있는 방법을 서술한다. 그리고 본 논문의 비교 실험에 쓰일 문서 분류 방법에 대한 관련 연구 및 현재 동향 대해 살펴본다.

1. 키워드 추출 방법

키워드 추출(Keyword Extraction)은 자연어 처리 분야에 속하는 연구로써, 문서에 나타나는 단어 중 핵심이 되는 단어를 추출하는 작업이다. 학습문서[8]에서 키워드를 구축 시 많은 단어를 키워드로 추출할 경우 효율이 떨어진다. 이 때문에 키워드 효율 저하 없이 키워드의 숫자를 줄이기 위한 연구가 진행되고 있다. 이와 관련하여 본 절에서는 문서의 키워드 추출과 관련된 대표적인 기법과 관련 연구를 살펴본다.

---

6) 텍스트 마이닝(Text mining)은 데이터 마이닝(Data mining)에 속하며, 저장된 대규모 데이터 안에서 반복되는 패턴이나 통계적 규칙을 찾아 분석하는 분야이다.. [10][11]
7) 텍스트 분석에서 해당 문서를 분석하기 위해 명사를 추출하는 작업을 말한다.
8) 기계 학습 안에서 훈련 데이터(Training Data)에 해당하는 문서를 말한다[12]

1) 이진 가중치(Boolean Weighting)

이진 가중치는 가중치 측정 방법 중 가장 단순한 방법이며, 정보검색 분야의 불리언 모델(Boolean Model)이다. 따라서 문서에 단어가 출현하면 1, 출현하지 않으면 0으로 표현하여 AND, OR, NOT 연산을 이용한다.

$$W_t = \begin{cases} 1 & \text{if } frequency_{(t)} > 0 \\ 0 & otherwise \end{cases} \quad \text{(수식 1)}$$

수식 1에서 $frequency_{(t)}$은 문서에 출현한 단어 t의 빈도이다. 이진 가중치는 단순하고 계산량이 적은 장점이 있어 사용자 질의 처리를 위한 단순 방법을 제공한다[13]. 예를 들어 윈도우 운영체제에는 포함되지만, 리눅스에는 포함되지 않는 질의어를 손쉽게 생성할 수 있다. 하지만 문서간의 유사도를 구할 수 없기 때문에 문서정보검색 분야에서 거의 사용되지 않는다.

2) 단어 빈도(Term Frequency)

문서에서 출현하는 단어의 개수를 그대로 나타낸 방법이다. 다음 절에 나올 TF-IDF, TF-ICF(TF-Inverse Category Frequency), TF-ISF(TF-Inverse Sentence Frequency) 방법도 기본 단어 빈도에 문서, 카테고리, 문장의 관계를 고려하여 가중치를 측정한 방법이다. 기본적으로 TF 값에 영향을 많이 받으며, 동일한 TF 조건일 경우 IDF, ICF, ISF의 관계에 따라 가중치 순위가 정해진다.

$$W_t = frequency_{(t)} \quad \text{(수식 2)}$$

3) TF-IDF

TF-IDF는 문서와 문서 사이에서 단어의 가중치를 측정하는 대표적인 방법으로 사용되고 있다. 단어 빈도(Term Frequency)를 역 문서 빈도(Inverse Document Frequency)와의 곱으로 나타내며, 단어를 포함하고 있는 문서의 빈도를 고려해 가중치를 측정하는 방법이다[14][15].

$$W_{(t,d)} = tf_{(t,d)} \times idf_{(t)} \quad \text{(수식 3)}$$

t는 단어를 나타내며, d는 문서를 나타낸다. 따라서 $W_{(t,d)}$는 문서 d 안에 있는 단어 t의 가중치(Weight)를 나타낸다. $tf_{(t,d)}$도 마찬가지로 문서 d 안에 있는 단어 t의 빈도를 나타낸다. $idf_{(t)}$는 단어 t가 출현하는 문서 빈도에 역을 취하며, 단어 t가 출현하는 문서가 많을수록 흔하게 쓰이는 단어이므로 가중치 값이 적게 적용된다. 역 문서 빈도는 수식 4과 같다.

$$idf_{(t)} = \log(\frac{N}{df_{(t)}}) \quad \text{(수식 4)}$$

N은 전체 문서 개수이며, $df_{(t)}$는 단어 t가 등장하는 문서의 빈도이다. 전체문서 개수를 문서 빈도로 나누어 로그를 취한다. $idf_{(t)}$ 값은 로그를 취하여 값이 적게 나오기 때문에 TF-IDF의 최종 결과 값은 $tf_{(t,d)}$의 영향을 많이 받는다.

유은순은 소설 텍스트 구조에서 소설의 주제 내용을 잘 나타내고 있는 머리말, 맺음말과 등장인물의 대사가 작가의 작품세계 및 주제를 나타내는 유용한 점을 고려한 대화문, 비대화 문으로 구조화한 후, 주제 단어에 TF-IDF 가중치를 부여하여 주제어를 추출하였다. 이 연구에서 본문 텍스트만 사용하여 가중치를 부여한 결과보다 대화문에 높은 가중치를 적용하고 머리말, 맺음말을 추가로 포함하였을 때 주제어를 추출한 정확도가 42.1% 향상되었다[16].

박호식은 의료 소견 데이터에서 TF-IDF 가중치를 적용해 문서 중 빈번하지 않

은 항목을 고려한 방법으로 문서마다 연관규칙을 적용해 질병으로 의심되는 후보 질병을 추론하였다. 또한, 추론된 결과를 의료 온톨로지를 이용하여 병명 간의 온톨로지 상의 거리를 구해 표현함으로써 의사결정에 시각적인 도움을 주었다[17].

4) TF-ICF

문서의 범주화에 특화된 방법으로, 역 범주 빈도(Inverted Category Frequency)는 소수의 범주에 많이 나오는 단어에 대해서 가중치를 높게 주고, 여러 범주에 자주 나오는 단어는 가중치를 낮게 부여하는 방법이다[15].

$$W_{(t,d)} = tf_{(t,d)} \times icf_{(t)} \quad \text{(수식 5)}$$

주로 대량의 문서로 범주화가 된 문서집단에서의 단어 가중치를 측정할 때 사용한다.

$$icf_{(t)} = \log(M) - \log(cf_{(t)}) + 1 \quad \text{(수식 6)}$$

수식 6에서 M은 전체 범주의 개수이며, $cf_{(t)}$는 단어 t가 등장하는 범주의 빈도이다. 전체 범주의 개수인 M의 빈도만큼 단어 t가 등장할 경우 $icf_{(t)}$ 값이 0이 되기 때문에 끝에 1을 더하였다.

이재욱은 한겨레신문의 각 카테고리의 단어에 TF-ICF 값을 적용하여 수치화한 후, 임계치 값을 이용해 단어를 추출하였고, 추출된 단어로 학습문서를 만들었다. 이후 문서 간의 정보량을 이용해 유사도를 측정하는 상호 정보량과 각 카테고리의 학습단어 개수가 달라서 학습빈도가 높았던 카테고리로 문서가 오분류 되었던 문제를 개선한 로그 정규화 방법을 적용하여 뉴스의 카테고리 오분류 문제를 개선하였다[18].

5) TF-ISF

TF-ISF는 문서 요약에 가장 많이 사용하는 방법이다. 역 문장 빈도(Inverse Sentence Frequency)는 특정 단어가 여러 문장에 많이 등장할수록 특정 단어의 가중치를 낮게 측정하는 방식이다[15][19].

$$W_{(t,d)} = tf_{(t,d)} \times isf_{(t)} \quad \text{(수식 7)}$$

TF-ISF는 문장을 이용하여 단어들의 가중치를 측정하기 때문에, 단일 문서에서 키워드를 추출할 때 사용된다.

$$isf_{(t)} = \log(\frac{|s|}{sf_{(t)}}) \quad \text{(수식 8)}$$

$sf_{(t)}$는 단어 t가 등장하는 문장의 개수이며, $|s|$는 전체 문장의 개수이다.

2. 문서 유사도 측정

문서 유사도 측정은 문서 간의 연관 관계를 측정하는 방법으로 각 문서 안의 대표 키워드를 추출 후, 문서들의 대표 키워드를 일치 여부로 유사도를 측정한다. 초기에는 문서에 나오는 단어의 단순 중복 빈도로 문서끼리의 유사도를 측정했다. 이후 문서와 문서의 연관 관계, 문서가 포함하고 있는 단어의 범주, 특정 문서의 구조에 따른 키워드의 중요도를 고려한 유사도 측정방법이 연구되고 있다. 문서 유사도 측정의 대표적 방법으로는 코사인 유사도, 문서 빈도, 카이제곱 통계량, 상호 정보량, 정보 획득량 등이 있다.

1) 코사인 유사도(Cosine Similarity)

단어 정보를 벡터로 표현하여 벡터의 크기인 빈도를 고려하지 않고 오로지 벡터의 방향만을 이용해 유사도를 측정하는 방법이 코사인 유사도다. 두 벡터 사이의 내각을 통해 각이 좁을수록 유사한 값을 가지며, 각이 넓을수록 유사하지 않은 값을 가진다. 따라서 코사인 각도가 $\cos 0^o$인 1 값과 $\cos 180^o$인 -1 값으로 표현된다.

$$\cos(A,B) = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad \text{(수식 9)}$$

코사인 유사도를 이용해 문서의 단어 유사도를 측정할 경우 단어가 음수 값이 되는 것은 불가능하므로 두 문서의 유사도는 0부터 1까지의 값으로 표현된다[20].

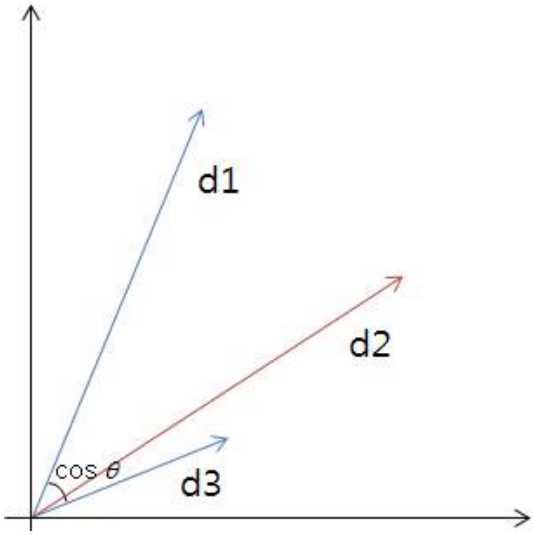


그림 3. 문서의 코사인 유사도 예시

그림 2에서 문서 d1, d2, d3는 각각 벡터로 표현되었다. d2와 가장 가까운 문서는 $\cos\theta$의 내각이 가장 작은 d3 문서이며, d1의 경우 벡터의 크기는 비슷하지만, 코사인 유사도는 벡터의 크기를 고려하지 않으므로 d3가 d1보다 d2에 더 유사한 값을 가진다.

2) 문서 빈도(Document Frequency)

단어가 나타난 문서의 빈도를 측정하는 단순한 방법이다. 학습문서 구축 시 단어가 출현하는 문서의 개수를 파악 후, 문서 안의 단어가 임계치[9] 조건에 만족할 경우 대표 단어로 추출한다. 이후 추출된 대표단어로 문서 간의 유사도를 측정한다. 이 기법의 경우 간단하고 계산량이 적은 이점이 있지만, 문서에서 출현 개수가 적은 단어는 추출하지 못하는 단점이 있다. 이런 이유로 저 빈도 단어의 정보량이 많다는 기본 가정에 어긋나 사용되지 않는다[21].

9) 특정 기준 조건을 충족하는 값.

3) 카이제곱 통계량(X^2 Statistics)

카이제곱 통계량은 단어와 범주의 의존성을 측정하는 방법이다[22]. 예를 들어 유권자의 성별이 구분된 투표 결과의 이원분류표[10]가 있을 때, 카이제곱 통계량을 이용하여 투표결과가 유권자의 성별과 상관이 있는지 또는 투표와 성별 간에 연관성 여부를 파악할 수 있다.

$$x^2(t,s) = \frac{M \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad \text{(수식 10)}$$

t: 단어, s: 카테고리, M: 전체 문서 개수
A: 카테고리 s에 속한 문서 중 단어 t를 포함한 문서 개수
B: 카테고리 s에 속하지 않은 문서 중 단어 t를 포함한 문서 개수
C: 카테고리 s에 속한 문서 중 단어 t를 포함하지 않은 문서 개수
D: 카테고리 s에 속하지 않은 문서 중 단어 t를 포함하지 않은 문서 개수

카이제곱 통계량은 수식 10으로 나타낼 수 있으며, 단어 t와 카테고리 s가 독립적이면 통계량은 0의 값을 가진다.

김진상은 카이제곱 통계량을 기반으로 알고리즘을 개선하고, 이를 이용해 스팸 메일의 빈도수뿐만 아니라 단어의 출현 빈도수를 반영하여 단어의 기본형으로 스태밍 한 후, 스팸 메일에 임계 값 이상으로 나타나는 단어를 제거하였다. 결과적으로 스팸 메일 필터링 방법과 더불어 정확률과 재현율을 동시에 향상 시키는 방법을 제안했다[24].

10) 측정하고자 하는 데이터의 상호관계를 행과 열의 이원적 관계로 나타낸 표[23].

4) 상호 정보량(Mutual Information)

통계적 언어 모델(Statistical language model)에서 주로 사용하는 기법이며[25], 두 문서가 공유하는 정보량의 크기로 유사도를 구하는 방식이다.

$$MI(A,B) = \log \frac{P(A \wedge B)}{P(A) \times P(B)} \quad \text{(수식 11)}$$

상호 정보량은 수식 11과 같다. 확률(Probability)을 사용하여 사건 A와 B의 정보량을 구하며, 사건 A와 B가 동시에 일어날 확률이 높을수록 상호 정보량이 크다. 각 사건이 일어날 확률값은 수식 12로 구할 수 있다.

$$P(K) = \frac{K\text{의 경우의 수}}{\text{전체 경우의 수}} \quad \text{(수식 12)}$$

확률을 기반으로 공유하는 정보를 계산하기 때문에 문서 안의 불필요 정보가 많으면 많을수록 확률값이 낮아지는 단점이 있다.

허정은 상호정보량과 복합명사 의미사전을 이용해 동음이의어의 의미 중의성 문제를 개선하였다. 이 연구에서 어휘 단어 간의 연관계수로 상호정보량을 이용하여 기존의 사전 뜻풀이를 이용해 단어들이 정확히 매칭되지 않았던 문제를 해결하였다. 또한, 상호정보량이 가지고 있는 어휘 쌍의 비율에 대한 가중치와 의미별 비율 가중치, 뜻풀이의 길이 가중치를 이용해 언어적 특징을 반영하였다. 그리고 복합명사를 이루는 단일명사들은 서로의 의미를 제약한다는 특징을 이용하여 빈도가 높은 복합명사를 추출해 의미사전을 구축하였고, 구축된 의미사전을 이용해 동음이의어 중의성 해소에 적용하였다[26].

5) 정보 획득량(Information Gain)

기계 학습 분야에서 주로 사용되는 기법으로, 문서의 출현 빈도와 출현하지 않은 빈도까지 고려하여 각 범주의 키워드 값을 계산한다. 각 범주의 집합을 $\{x_{1,}x_{2,}\cdots_{,}x_{n}\}$이라 할 때 수식 13으로 구할 수 있다.

$$G(t) = -\sum_{i=1}^{m} P(x_i)P(x_i) + P(t)\sum_{i=1}^{m} P(x_i|t)\log P(x_i|t) + P(t)\sum_{i=1}^{m} P(x_i|\bar{t})\log P(x_i|\bar{t}) \quad \text{(수식 13)}$$

모든 범주의 평균값으로 계산되며, 확률을 이용해 범주에 속한 문서의 단어 t에 대한 정보 획득량을 계산하여 임계치 값을 만족하는 단어를 추출하고 학습데이터를 구성한다. 정보 획득량은 문서 내 특정 키워드의 출현 여부로써 문서의 정보를 측정한다. 이 때문에 전체 출현 빈도가 작은 키워드에 큰 정보량을 부여할 수 있다는 장점이 있다.

Ⅲ. 문서 내 단어 분포를 고려한 키워드 가중치 측정

기존의 키워드 추출은 문서 안의 단어 빈도에 따라 빈도가 높은 단어가 해당 문서의 대표 키워드로 추출되었다. 하지만 특정 장르의 문서에서 높은 빈도를 가진 단어가 다른 장르의 문서들에서도 높게 나타날 경우, 장르 분류에 있어 변별력을 가지지 못하는 문제가 있었다. 이러한 문제를 해결하기 위해 본 논문에서는 각 문서 내 단어 분포를 측정해 분포도가 평균 이상인 단어에 대해 추가 가중치를 부여하여 문서의 키워드를 추출하는 방법을 제안한다. 먼저 신조어 및 줄임말로 단어 간의 형태 매칭이 어려운 키워드에 대해 위키피디아의 동의어 문서 정보를 이용해 대표 단어를 선정하고, 제안하는 방법으로 문서 간의 단어 가중치를 측정한다. 그림 3는 본 논문의 전체 흐름도를 나타낸다.

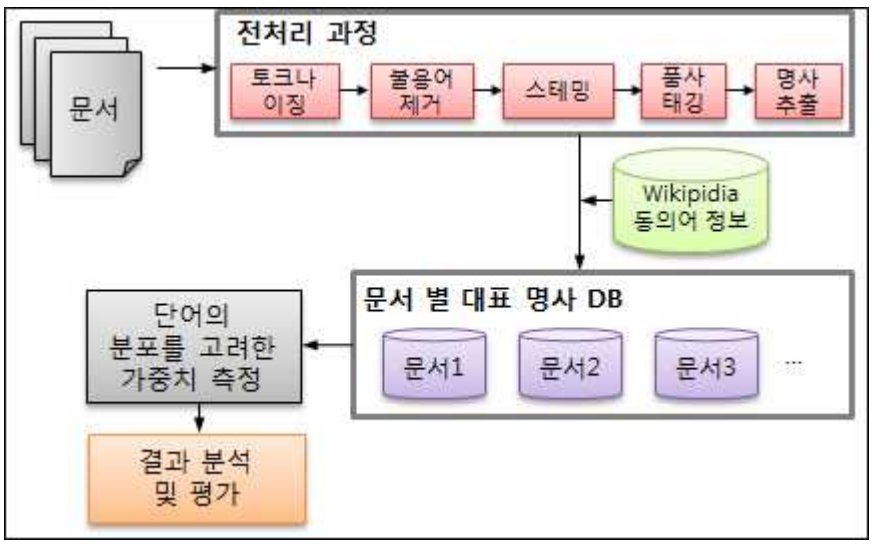


그림 3. 전체 흐름도

1. 전처리(Preprocessing)

전처리는 자연어 처리 분야에서 글이나 문서를 분석하기 위한 기초 작업이다. 영어문서에서 문장을 띄어쓰기 단위로 토큰화시킨 후, 불용어 사전을 통해 불용어를 제거한다. 이후 과거형, 미래형, 복수형 등의 단어에 대해 원형으로 바꾸는 스태밍 과정을 거쳐 사전을 통해 품사를 판별하고 품사 태깅을 한다. 본 논문에서는 품사 태깅된 단어 중에서 문서를 효율적으로 표현할 수 있는 명사를 추출한다. 본 절에서는 전처리 과정을 자세히 서술한다.

1) 토크나이징(Tokenizing)

토크나이징은 문장을 하나의 단어 단위로 세분화시키는 작업이다. 문장에서 띄어쓰기를 기준으로 단어를 나누며, 전처리 과정의 초기 과정이다.

2) 불용어(Stop word) 제거

불용어 제거는 문장에서 불필요한 단어를 제거하는 작업으로 접속사, 전치사, 관사, 조사 등 의미가 없는 단어를 불용어 사전을 이용해 제거한다. 표 2[11]는 워드넷(WordNet)에서 사용하는 불용어 사전이다. 불필요한 단어를 사전에 제거함으로써 스태밍, 품사 태깅 과정에서 좀 더 효율적으로 작업을 처리할 수 있다.

11) A WordNet Stop List[27]

표 2. 불용어 목록 예시

	불용어		불용어		불용어
1	I	13	against	25	everybody
2	a	14	amid	26	everyone
3	an	15	amidst	27	for
4	as	16	among	28	from
5	at	17	amongst	29	her
6	by	18	and	30	hers
7	he	19	anybody	31	herself
8	his	20	anyone	32	him
9	me	21	because	33	himself
10	or	22	beside	34	hisself
11	thou	23	circa	35	idem
12	who	24	during	...	...

3) 스태밍(Stemming)

스태밍은 단어를 원형의 형태로 변환시키는 작업이다. 영어의 문법은 주어와 동사의 단수형, 복수형에 따른 수일치 때문에 단어에 -s, -es를 붙이며, 동사의 시제는 과거, 현재, 미래 등으로 나뉜다. 이러한 이유 때문에 통일성을 위해 단어의 원형으로 변환시킨다. 표 3은 입력문장을 스태밍 작업한 예시이다.

표 3. 스태밍 작업 예시

	예시
문장[12]	Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for price cut, volume of sales.
스태밍	document, describe, marketing, strategy, carry, company, agriculture, chemical, report, prediction, market, share, chemical, report, market, statistic, price, cut, volume, sale

12) Search Engines 저서에의 4.3장 Document Parsing의 Stemming 문장[28].

4) 품사 태깅(POS Tagging)

품사 태깅은 스태밍 된 단어의 품사를 판별하는 작업이다. 본 연구에서는 자연어처리 연구소인 샤인웨어[13]에서 자바로 만든 영어 형태소 분석기인 EN-POSTA 0.5 버전을 사용하였으며, 영어 문장에서 품사 태깅에 사용되는 품사 규칙 기호[14]는 표 4와 같다.

표 4. 영어 단어의 품사기호

기호	의미	기호	의미	기호	의미
DT	한정사	JJS	형용사, 최상급	CC	등위 접속사
QT	수량사	JJR	형용사, 비교급	UH	감탄사
CD	기수	JJ	형용사	RP	소사
NN	명사, 단수형	MD	법조동사	SYM	기호
NNS	명사, 복수형	VB	동사, 기본형	$	통화 기호
NNP	고유명사, 단수형	VBP	동사, 현재형, 3인칭 단수 제외	“	큰/작은 따음표
NNPS	고유명사, 복수형	VBZ	동사, 현재형, 3인칭 단수	(	여는 괄호
TO	전치사	VBD	동사, 과거형	)	닫는 괄호
PRP	인칭대명사	VBN	동사, 과거분사	,	쉼표
PRP$	소유대명사	VBG	동사, 동명사, 현재분사	.	구문 끝 구두점
POS	소유 종료	IN	전치사, 종속 접속사	:	구문 중간 구두점
RBS	부사, 최상급	RBR	부사, 비교급	RB	부사
UNKNOWN	알 수 없는 단어				
WRB	I like you when you eat something 구문의 When 같은 형용사				
WP	관계대명사로 사용될 때의 which 및 that 같은 대명사				
WP$	소유격 대명사 (예: whose)				
WDT	Which book do you lkie better 구문의 which 같은 한정사				
EX	There was a party 구문에서와 같이, 존재 there				

13) 자연어 처리 연구소[29]
14) 영어 문서에 사용하는 품사 태그[30]

표 5는 품사 태깅 알고리즘을 나타낸다.

표 5 품사 태깅 알고리즘

```
import java.io.BufferedReader;
import java.io.FileReader;
import java.util.List;

import kr.co.shineware.nlp.posta.en.core.EnPosta;

public class n_extraction {
        public static void main(String[] args) throws Exception{
                //형태소 분석 라이브러리
                EnPosta posta = new EnPosta();
                posta.load("C:₩₩models₩₩model_0.5");
                posta.buildFailLink();
                //테스트 파일
                FileReader fr = new FileReader("documnet1.txt");
                BufferedReader br = new BufferedReader(fr);

                String fl = null;
                do{
                        fl=br.readLine();
                        List<String> resultList = posta.analyze(fl);
                        for (String result : resultList) {
                //태깅 결과
                                System.out.println(result);
                                }
                        }while(!(fl==null));
                br.close();
        }
}
```

표 6과 같이 입력 문장을 품사 태깅 알고리즘에 입력하면, 단어의 품사를 얻을 수 있다.

표 6. 입력 문장을 대상으로 품사 태깅한 예시

	예 시
입력문장	Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for price cut, volume of sales.
품사 태깅 결과	Document/NNP, will/MD, describe/VB, marketing/NN, strategies/NNS, carried/VBD, out/RP, by/IN, U.S/NN, companies/NNS, for/IN, their/PRP$, agricultural/JJ, chemicals/NNS, report/NN, predictions/NNS, for/IN, market/NN, share/NN, of/IN, such/JJ, chemicals/NNS, or/CC, report/NN, market/NN, statistics/NNS, for/IN, price/NN, cut/NN, volume/NN, of/IN, sales/NNS

5) 명사 추출

품사 태깅을 통해 얻은 품사 중 NN 기호가 포함된 단어를 명사로 추출한다.

2. 위키피디아를 이용한 단어 확장 방법

위키피디아는 온라인 백과사전으로 다수의 사용자가 정보 제공자로 참여하며, 자유로운 콘텐츠(contents)와 다국적 언어로 광범위한 정보를 공유하고 있다. 또한, 위키피디아는 실시간으로 정보가 생성되고, 갱신되기 때문에 워드넷[15]과는 다르게 시간 경과에 대한 문서 정보의 유효성 제약을 받지 않는다. 그리고 전문적 사전, 뉴스 기사, 서적, 연구 문헌, 공식적 자료 등을 기반으로 문서가 기재되기 때문에 문서의 정보와 개념에 대한 신뢰성을 가지고 있다[32].

본 논문에서는 위키피디아에서 제공하는 문서 정보인 redirect를 이용하여 신조어, 줄임말, 동의어 등의 단어를 하나의 대표 단어로 추출방법을 서술한다.

1) 위키피디아를 이용한 정보 추출

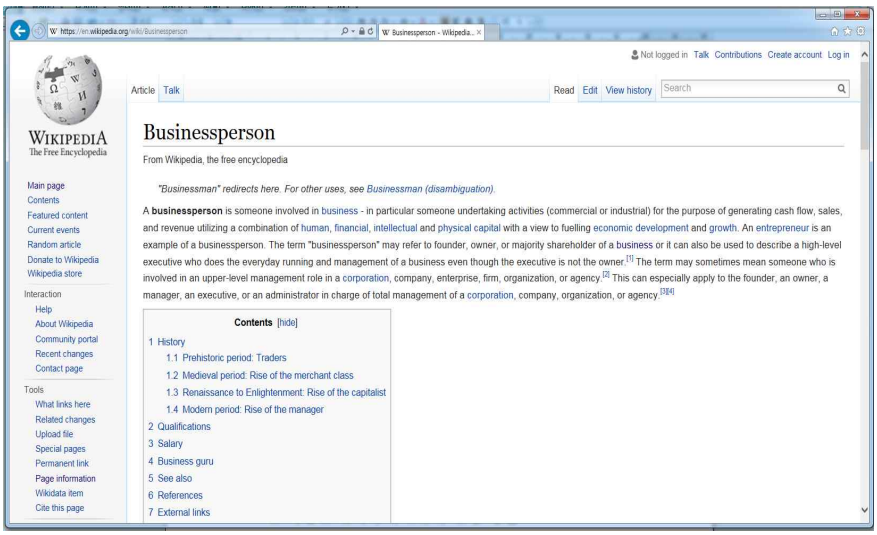


그림 5. 위키피디아 문서 정보 예시(Businessperson)

15) 영어의 의미 어휘목록[31].

사용자들이 많이 쓰는 단어일수록 문서 정보를 많이 이용하게 되고, 단어와 관련된 링크나 정보를 생성하게 된다. 생성된 정보는 다시 문서 정보로 축적이 되어 여러 문서와의 관련 링크나, 카테고리로 분류가 된다. 또한, 위키피디아는 문서에 대한 정의문과 목차를 나누어 상세하게 설명하고 있으므로 전문적인 정보나 관련된 링크 정보를 쉽게 얻을 수 있다. 그림 4은 위키피디아 백과사전의 Businessperson 단어의 문서 정보를 나타내고 있다. 왼쪽의 Tools 기능을 통하여 문서 정보, 링크 정보 등을 열람할 수 있다.

2) Redirect 기능을 이용한 대표 단어 추출

위키피디아는 Redirect 기능을 이용하여 신조어, 줄임말, 동의어 등 의미가 같은 단어를 하나의 대표 단어로 나타내고 있다. 이 때문에 문서 안에서 의미가 같지만, 형태가 달라 서로 다른 가중치 점수를 부여했던 문제점을 개선할 수 있다. 표 7은 Redirect에 해당하는 단어 중 일부분이다.

표 7. 위키피디아의 Redirect 단어

단어	Redirected word
Businessperson	Businessman, Businesspeople, Businesswoman
Cadaver	Corpse, Cadavar, Carcase
Child	Children, Schoolchild, Sproggen, Kiddies
Ceramic	Keramika, Pottery
Event	Occasion
Forever	4ever
Injustice	Unfairness, Wrength, Unjust
Law	Illegality, Legal,
League of Legends	LOL, Dominion, Runeterra
Mathematics	Math, Mathmetics, Matheamtics, Matemathics
Professor	Prof, Professora, Professorship, Catedratico
Promotion	Promo
Rebelion	Insurrection, Rebels, Uprising, Insurrectionary
Test	Examinations, Tester

각기 다른 문서에 Businessman, Businesspeople, Businesswoman 단어가 출현했을 경우, 같은 의미이지만 형태가 달라서 서로 다른 단어로 구분되었다. 이러한 단어들을 대표 단어인 Businessperson으로 변환하여 문서 간의 단어 중복 빈도를 높이면, 같은 부류에 해당하는 문서들의 유사도를 더 높일 수 있다.

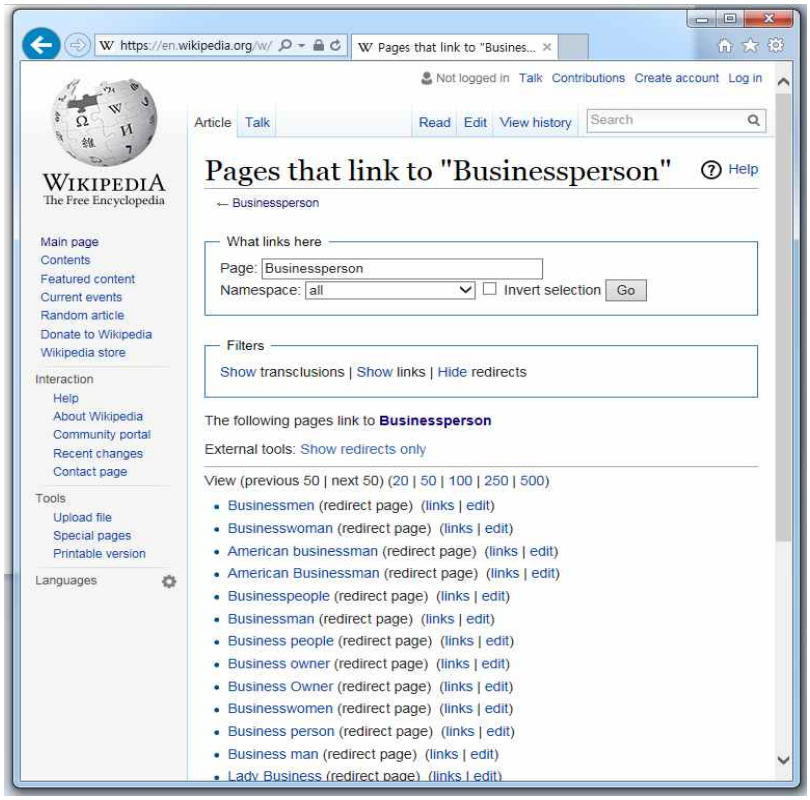


그림 6. Businessperson 단어의 redirect 정보

3. 단어의 분포를 이용한 가중치 측정 방법

위키피디아의 Redirect 정보를 이용하여 신조어, 줄임말, 동의어를 대표 단어로 변형시키고 문서별로 단어 데이터베이스를 구축한다. 구축된 단어 데이터베이스를 이용해 문서마다 TF-IDF 값으로 순위화한다. 각 문서의 순위화된 단어와 단어가 등장하지 않는 문서의 관계를 고려한 분포도 편차를 측정한다. 측정된 값은 순위가 평균 이상인 단어에만 추가해 가중치를 구한다.

$$\sqrt{\frac{(x_1-m)^2+(x_2-m)^2+\ \cdots\ +(x_n-m)^2}{n}} \qquad \text{(수식 14)}$$

단어들이 분포되어있는 값을 구하기 위해서는 표준편차를 사용한다. 일반적인 표준편차는 자료의 평균을 기준으로 하였을 때, 자료가 평균 주변에 모여있는지 아니면 흩어져있는지를 구할 수 있으며, 수식 14와 같다. 각 자료의 값은 $x_1, x_2, \cdots, x_n$을 나타내며, 자료들의 평균값은 m이다. 일반적인 표준편차는 해당 단어가 출현할 경우에만 그 값을 포함하여 측정한다.

본 논문에서는 단어가 출현하지 않은 문서에도 단어의 가중치 값을 측정하는 방법을 제시한다. 특정 단어가 출현한 문서와 출현하지 않은 문서의 특정 단어에 대해 편찻값을 크게 주기 위해 단어가 등장하지 않은 문서의 최저 순위에 1을 더하여 단어의 순위를 부여하고 편찻값을 측정하는 방법을 제안한다.

1) 단어 분포 편차가 큰 경우

그림 6는 TF-IDF 가중치를 적용하여 문서마다 가중치가 높은 순으로 순위를 매겨놓은 예시이다.

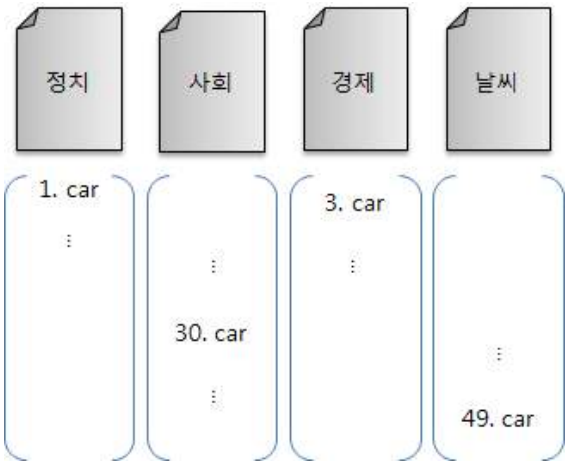


그림 7. 단어의 분포 편차가 큰 경우

그림 6에서 단어 car는 단어들이 분포된 순위의 편차가 크며, 각 단어의 순위 값을 이용해 표준편차를 구하면 표 8과 같다.

표 8. 단어 car의 표준 편차 값

n	4
m	$\frac{1+30+3+49}{4}=20.75$
수식 대입	$\sqrt{\frac{(1-20.75)^2+(30-20.75)^2+(3-20.75)^2+(49-20.75)^2}{4}}$
표준 편차	19.929

단어 car는 정치와 경제에 주요 단어지만, 사회와 날씨에는 분별력을 가지지 못한다. 이 때문에 단어의 평균을 임계점으로 20.75보다 높은 순위인 정치, 경제 car 단어의 TF-IDF에 추가 편찻값을 부여하였다.

2) 단어 분포 편차가 적은 경우

그림 7은 단어가 분포된 순위의 편차가 적은 예시이다.

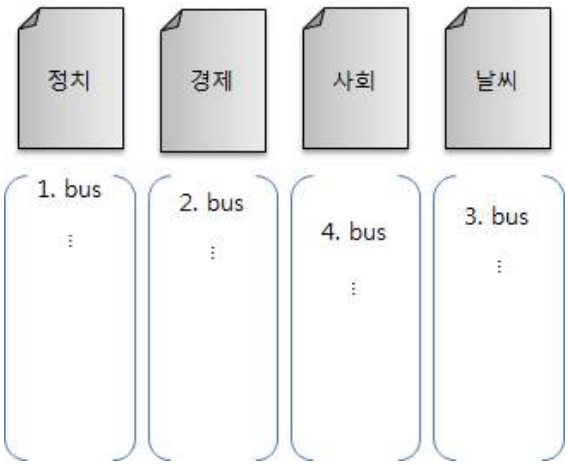


그림 8. 단어의 분포 편차가 적은 경우

단어 bus는 순위가 높아 모든 문서의 주요 키워드이지만, 각 문서를 분류하여 구분하기에는 분별력이 떨어진다.

표 9. 단어 bus의 표준 편차 값

n	4
m	$\frac{1+2+4+3}{4}=2.5$
수식 대입	$\sqrt{\frac{(1-2.5)^2+(2-2.5)^2+(4-2.5)^2+(3-2.5)^2}{4}}$
표준 편차	1.118

단어 bus와 같이 문서 분류에 분별력이 떨어지는 단어의 편찻값은 1.118로 적게 나타나지만, 문서 분류에 분별력이 있는 단어 car의 편찻값은 19.929로 높게 나타났다.

그림 8과 같이 단어의 평균보다 높은 단어에 가중치를 적용함으로써 정치, 경제 문서에 주요 단어인 car의 TF-IDF 값에 추가 가중치를 부여하였다. 또한, 문서 분류에 변별력이 없는 bus와 같은 단어는 낮은 가중치를 부여하였다.

그림 9. 분포가 큰 단어의 표준편차 가중치

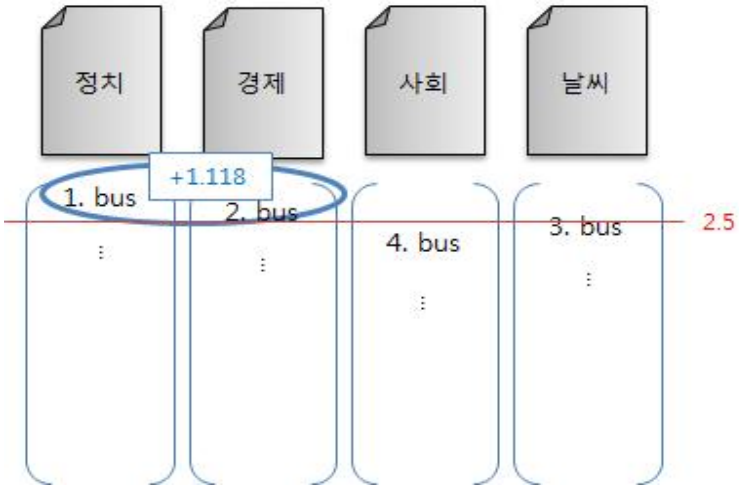


그림 10. 분포가 작은 단어의 표준편차 가중치

3) 모든 문서를 고려한 편차 측정 방법

표준편차는 해당 자료가 문서 내에 존재할 경우에 그 값을 이용하여 편찻값을 측정한다. 이 경우 모든 문서의 관계를 고려하지 못한다. 예를 들어 정치, 스포츠 문서가 있는데 단어 car가 정치 문서에만 나올 경우, 정치에만 국한된 표준편차 값을 얻는다.

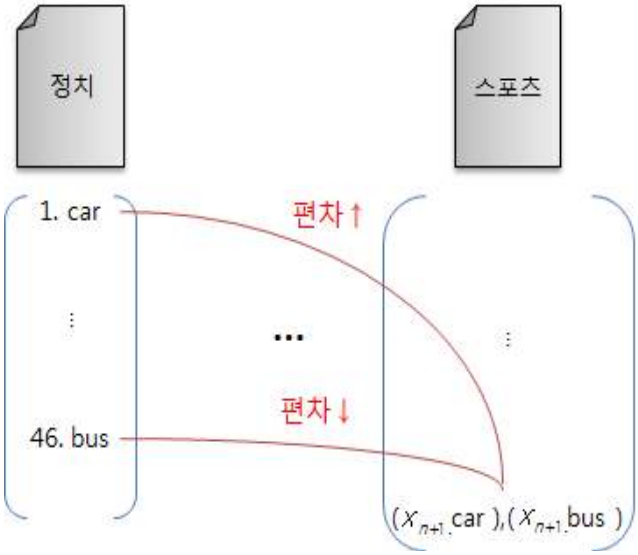


그림 11. 단어가 출현하지 않는 문서의 표준편차 측정 방법

본 논문에서는 이 문제를 해결하기 위해 스포츠 문서에도 가상의 순위화 된 단어를 넣는다. 스포츠 문서의 최저순위 다음으로 순위를 넣음으로써 그림 10과 같이 정치 문서의 주요 단어인 car는 높은 편찻값을 분별력이 없는 bus 단어에는 낮은 편찻값이 적용되었다.

제안하는 편차 적용 방법은 표 10과 같다.

표 10. 단어의 편차 값 적용 방법

	수 식
편차 가중치	$SD_i = \sqrt{\frac{\sum_{i,j=1}^{i=k,j=n}(x_{ij}-m_i)^2}{n^2}}$
최종 가중치	$W_{ij} = TF_{ij} \times (IDF_i + \log(SD_i))$
설명	$k = frequency_{(x)}$ (모든 단어의 개수) $n = frequency_{(document)}$ (문서의 개수) x_{ij} = 문서j에서 단어i의 순위값 m_i = 단어i의 평균

본 연구에서 제안하는 방법은 줄임말, 신조어, 동의어를 위키피디아를 이용해 대표단어로 변환시키고 모든 문서를 고려한 단어의 분포를 이용해 편찻값을 측정한다. 기존 TF-IDF에서 분별력이 없었던 IDF에 편찻값을 추가로 더해 가중치를 측정함으로써 중요한 단어에 분별력을 가지게 하여 각 문서를 대표할 수 있는 주요 키워드들을 선별하고자 한다.

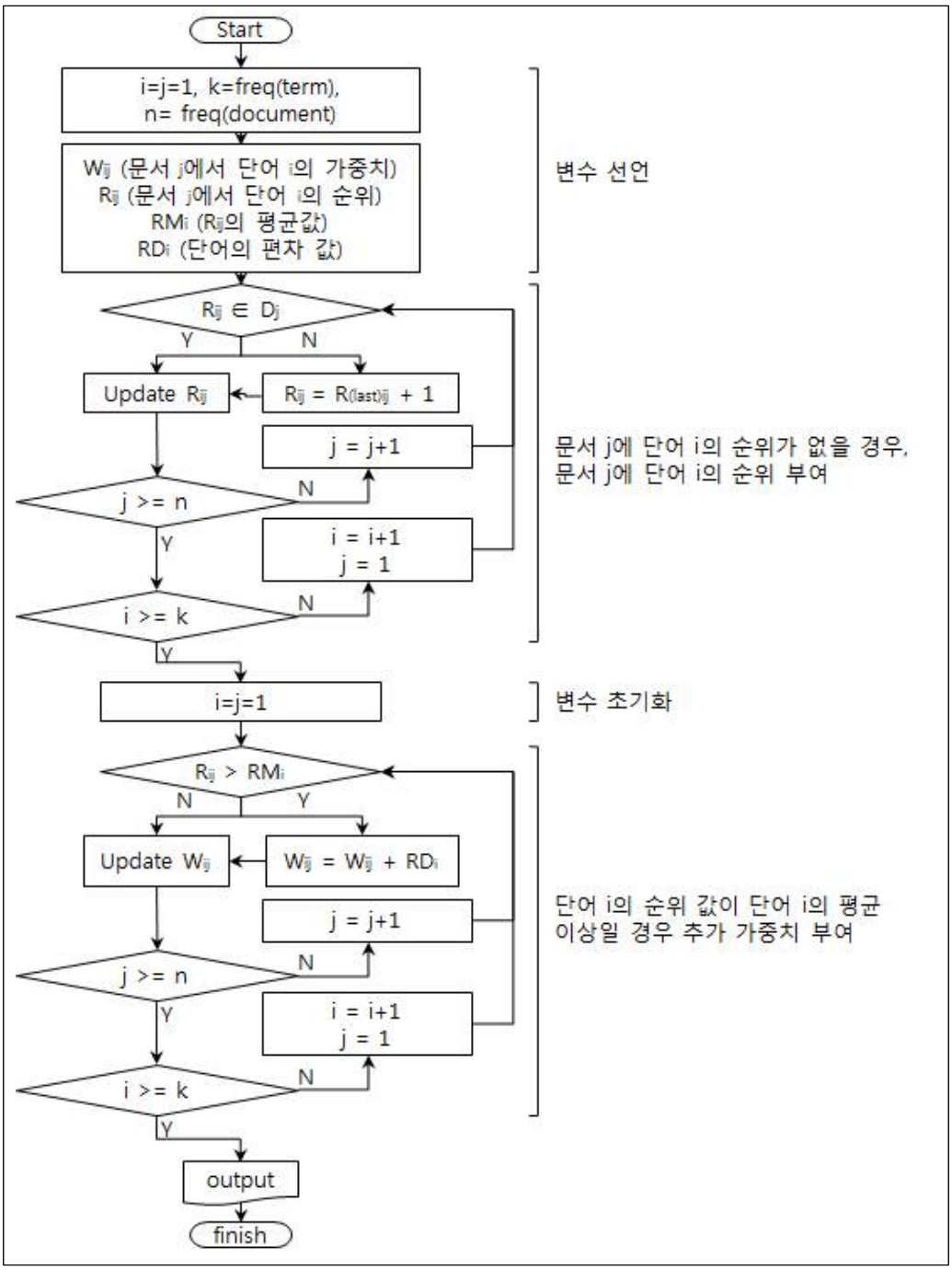
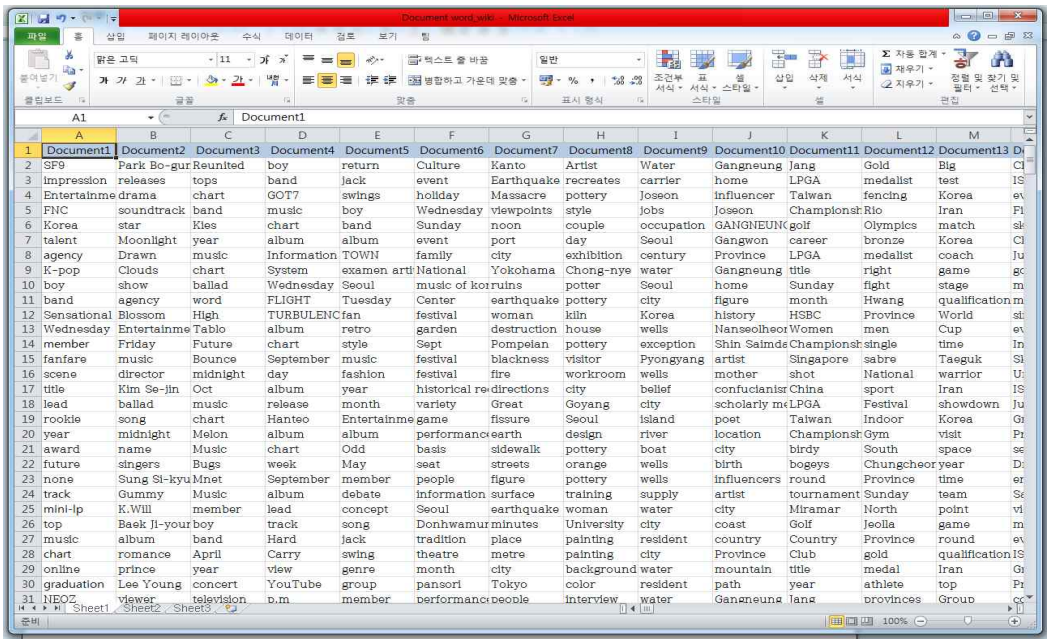


그림 11. 단어 순위의 분포 차를 고려한 가중치 적용 알고리즘

Ⅳ. 실험 및 평가

1. 실험 데이터

본 논문의 방법으로 측정된 키워드의 성능을 평가하기 위해, 키워드를 이용해 카테고리별 사전데이터를 구축하고, 구축된 사전데이터로 실험문서의 카테고리를 분류하는 실험을 하였다. 실험 데이터는 한국 타임스[16] 신문의 2014년 1월부터 2016년 9월까지 Game, Sport, Traditional, Entertainment 카테고리에 해당하는 문서 각 100개씩, 총 400 문서를 대상으로 구축하였으며, 실험문서는 2014년 1월부터 2016년 9월까지 각 500개씩 2000 문서를 대상으로 카테고리 분류 실험을 하였다.



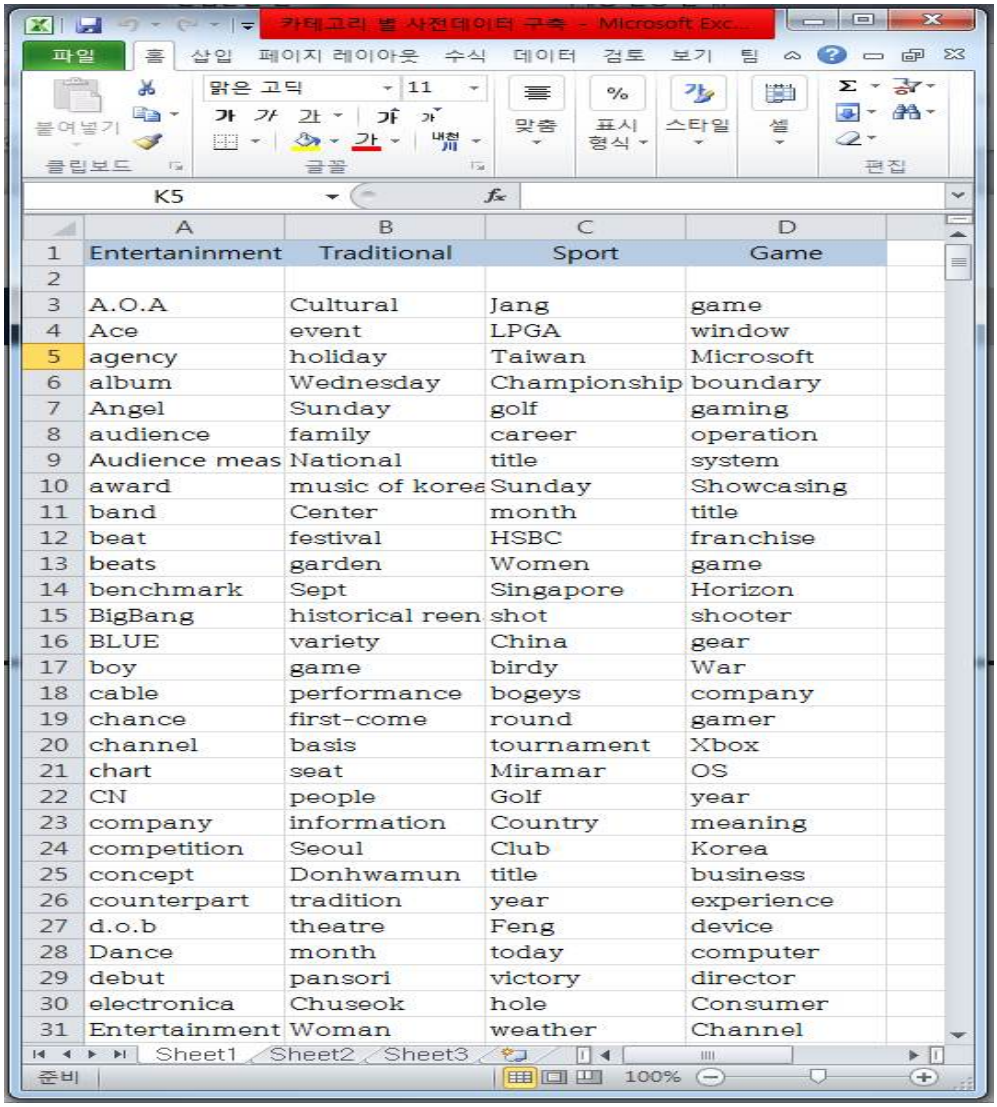
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Document1	Document2	Document3	Document4	Document5	Document6	Document7	Document8	Document9	Document10	Document11	Document12	Document13
2	SF9	Park Bo-gur	Reunited	boy	return	Culture	Kanto	Artist	Water	Gangneung	Jang	Gold	Big
3	Impression	releases	tops	band	jack	event	Earthquake	recreates	carrier	home	LPGA	medalist	test
4	Entertainme	drama	chart	GOT7	swings	holiday	Massacre	pottery	Joseon	influencer	Taiwan	fencing	Korea
5	FNC	soundtrack	band	music	boy	Wednesday	viewpoints	style	jobs	Joseon	Championsh	Rio	Iran
6	Korea	star	Kles	chart	band	Sunday	noon	couple	occupation	GANGNEUN(	golf	Olympics	match
7	talent	Moonlight	year	album	album	event	port	day	Seoul	Gangwon	career	bronze	Korea
8	agency	Drawn	music	Information	TOWN	family	city	exhibition	century	Province	LPGA	medalist	coach
9	K-pop	Clouds	chart	System	examen arti	National	Yokohama	Chong-nye	water	Gangneung	title	right	game
10	boy	show	ballad	Wednesday	Seoul	music of kor	ruins	potter	Seoul	home	Sunday	fight	stage
11	band	agency	word	FLIGHT	Tuesday	Center	earthquake	pottery	city	figure	month	Hwang	qualification
12	Sensational	Blossom	High	TURBULENC	fan	festival	woman	kiln	Korea	history	HSBC	Province	World
13	Wednesday	Entertainme	Tablo	album	retro	garden	destruction	house	wells	Nanseolheor	Women	men	Cup
14	member	Friday	Future	chart	style	Sept	Pompeian	pottery	exception	Shin Saimda	Championsh	single	time
15	fanfare	music	Bounce	September	music	festival	blackness	visitor	Pyongyang	artist	Singapore	sabre	Taeguk
16	scene	director	midnight	day	fashion	festival	fire	workroom	wells	mother	shot	National	warrior
17	title	Kim Se-jin	Oct	album	year	historical re	directions	city	belief	confucianisr	China	sport	Iran
18	lead	ballad	music	release	month	variety	Great	Goyang	city	scholarly m	LPGA	Festival	showdown
19	rookie	song	chart	Hanteo	Entertainme	game	fissure	Seoul	island	poet	Taiwan	Indoor	Korea
20	year	midnight	Melon	album	album	performanc	earth	design	river	location	Championsh	Gym	visit
21	award	name	Music	chart	Odd	basis	sidewalk	pottery	boat	city	birdy	South	space
22	future	singers	Bugs	week	May	seat	streets	orange	wells	birth	bogeys	Chungcheor	year
23	none	Sung Si-kyu	Mnet	September	member	people	figure	pottery	wells	influencers	round	Province	time
24	track	Gummy	Music	album	debate	information	surface	training	supply	artist	tournament	Sunday	team
25	mini-lp	K.Will	member	lead	concept	Seoul	earthquake	woman	water	city	Miramar	North	point
26	top	Baek Ji-your	boy	track	song	Donhwamur	minutes	University	city	coast	Golf	Jeolla	game
27	music	album	band	Hard	jack	tradition	place	painting	resident	country	Country	Province	round
28	chart	romance	April	Carry	swing	theatre	metre	painting	city	Province	Club	gold	qualification
29	online	prince	year	view	genre	month	city	background	water	mountain	title	medal	Iran
30	graduation	Lee Young	concert	YouTube	group	pansori	Tokyo	color	resident	path	year	athlete	top
31	NEOZ	viewer	television	p.m	member	performanc	people	interview	water	Gangneung	Jang	provinces	Group

그림 13. 한국 타임즈 신문에서 추출한 문서 집단 예

16) 한국 타임즈 신문(The Korea Times Subsection)[33].

2. 카테고리 별 사전데이터 구축

수집된 문서를 위키피디아를 이용해 신조어, 줄임말, 동의어를 대표단어로 변환하고 카테고리별 사전데이터를 구축하였다.

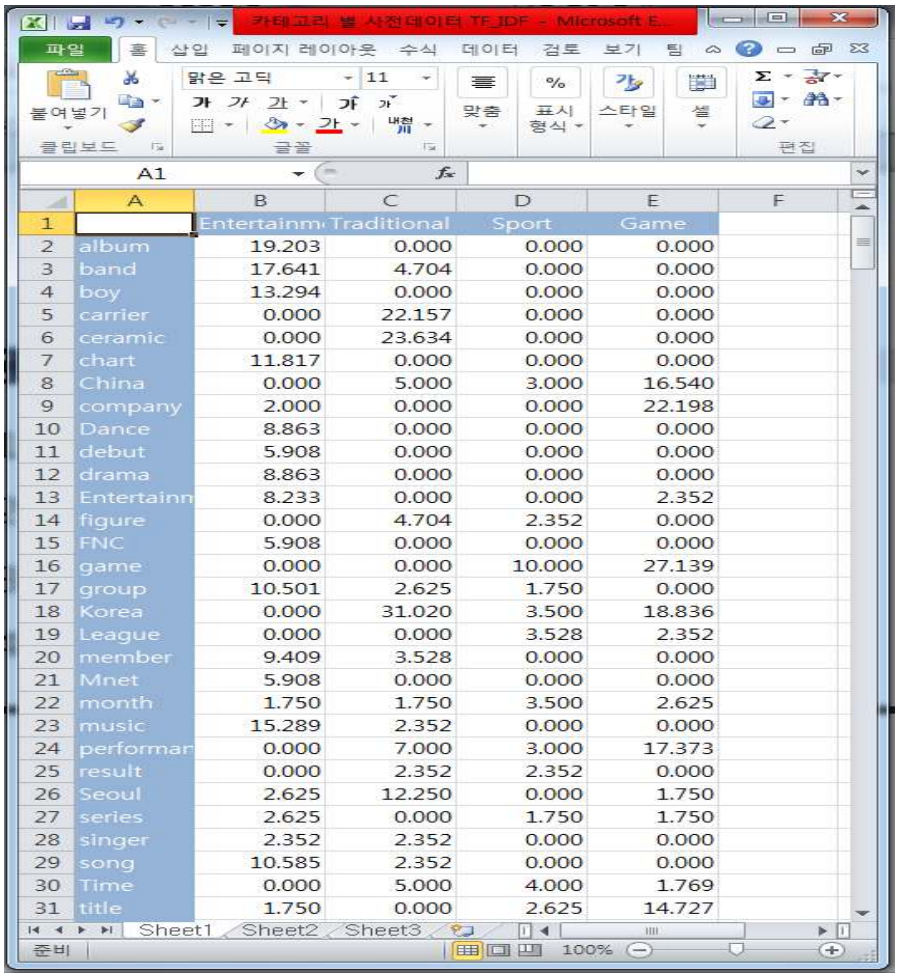


	A	B	C	D
1	Entertaninment	Traditional	Sport	Game
2				
3	A.O.A	Cultural	Jang	game
4	Ace	event	LPGA	window
5	agency	holiday	Taiwan	Microsoft
6	album	Wednesday	Championship	boundary
7	Angel	Sunday	golf	gaming
8	audience	family	career	operation
9	Audience meas	National	title	system
10	award	music of korea	Sunday	Showcasing
11	band	Center	month	title
12	beat	festival	HSBC	franchise
13	beats	garden	Women	game
14	benchmark	Sept	Singapore	Horizon
15	BigBang	historical reen	shot	shooter
16	BLUE	variety	China	gear
17	boy	game	birdy	War
18	cable	performance	bogeys	company
19	chance	first-come	round	gamer
20	channel	basis	tournament	Xbox
21	chart	seat	Miramar	OS
22	CN	people	Golf	year
23	company	information	Country	meaning
24	competition	Seoul	Club	Korea
25	concept	Donhwamun	title	business
26	counterpart	tradition	year	experience
27	d.o.b	theatre	Feng	device
28	Dance	month	today	computer
29	debut	pansori	victory	director
30	electronica	Chuseok	hole	Consumer
31	Entertainment	Woman	weather	Channel

그림 14. 위키피디아를 이용한 사전데이터 구축 예

3. 표준편차를 적용한 키워드 가중치 측정

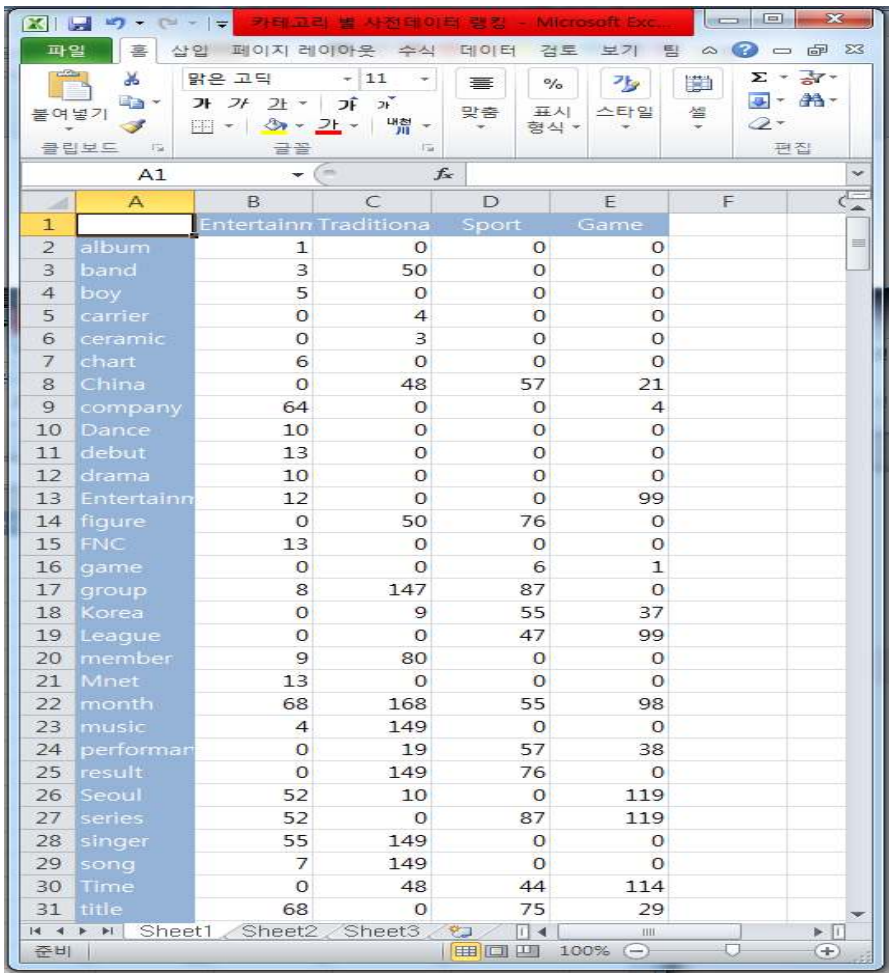
각 카테고리의 단어는 TF-IDF를 이용해 가중치를 측정하고, 카테고리마다 가중치를 구해 단어를 순위화한다. 그림 14은 각 카테고리 단어의 가중치 값을 나타낸다.



	A	B	C	D	E	F
1		Entertainm	Traditional	Sport	Game	
2	album	19.203	0.000	0.000	0.000	
3	band	17.641	4.704	0.000	0.000	
4	boy	13.294	0.000	0.000	0.000	
5	carrier	0.000	22.157	0.000	0.000	
6	ceramic	0.000	23.634	0.000	0.000	
7	chart	11.817	0.000	0.000	0.000	
8	China	0.000	5.000	3.000	16.540	
9	company	2.000	0.000	0.000	22.198	
10	Dance	8.863	0.000	0.000	0.000	
11	debut	5.908	0.000	0.000	0.000	
12	drama	8.863	0.000	0.000	0.000	
13	Entertainn	8.233	0.000	0.000	2.352	
14	figure	0.000	4.704	2.352	0.000	
15	FNC	5.908	0.000	0.000	0.000	
16	game	0.000	0.000	10.000	27.139	
17	group	10.501	2.625	1.750	0.000	
18	Korea	0.000	31.020	3.500	18.836	
19	League	0.000	0.000	3.528	2.352	
20	member	9.409	3.528	0.000	0.000	
21	Mnet	5.908	0.000	0.000	0.000	
22	month	1.750	1.750	3.500	2.625	
23	music	15.289	2.352	0.000	0.000	
24	performar	0.000	7.000	3.000	17.373	
25	result	0.000	2.352	2.352	0.000	
26	Seoul	2.625	12.250	0.000	1.750	
27	series	2.625	0.000	1.750	1.750	
28	singer	2.352	2.352	0.000	0.000	
29	song	10.585	2.352	0.000	0.000	
30	Time	0.000	5.000	4.000	1.769	
31	title	1.750	0.000	2.625	14.727	

그림 15. 카테고리별 단어 TF-IDF 값 예시

그림 15는 각 카테고리의 단어 순위를 나타내며, 카테고리에 단어가 등장하지 않을 경우 0 값을 나타낸다.



	A	B	C	D	E
1		Entertainm	Traditiona	Sport	Game
2	album	1	0	0	0
3	band	3	50	0	0
4	boy	5	0	0	0
5	carrier	0	4	0	0
6	ceramic	0	3	0	0
7	chart	6	0	0	0
8	China	0	48	57	21
9	company	64	0	0	4
10	Dance	10	0	0	0
11	debut	13	0	0	0
12	drama	10	0	0	0
13	Entertainm	12	0	0	99
14	figure	0	50	76	0
15	FNC	13	0	0	0
16	game	0	0	6	1
17	group	8	147	87	0
18	Korea	0	9	55	37
19	League	0	0	47	99
20	member	9	80	0	0
21	Mnet	13	0	0	0
22	month	68	168	55	98
23	music	4	149	0	0
24	performar	0	19	57	38
25	result	0	149	76	0
26	Seoul	52	10	0	119
27	series	52	0	87	119
28	singer	55	149	0	0
29	song	7	149	0	0
30	Time	0	48	44	114
31	title	68	0	75	29

그림 16. 카테고리 별 단어 순위화

단어를 모든 카테고리에 등장시켜 편차를 구하기 위해서 값이 0인 단어들은 각 문서의 최저순위보다 1순위를 더 낮추어 각 단어의 표준편차를 구한다.

표 11. Entertainment 카테고리 단어의 편차 값을 적용한 가중치 예

단어	TF	IDF	TF-IDF	log(편차)	**최종 가중치**
band	15	1.176	17.641	1.360	**38.047**
album	13	1.477	19.202	1.381	**37.151**
music	13	1.176	15.289	1.433	**33.922**
group	12	0.875	10.500	1.408	**27.392**
song	9	1.176	10.584	1.423	**23.394**
boy	9	1.477	13.294	1.365	**25.577**
chart	8	1.477	11.816	1.361	**22.702**
member	8	1.176	9.408	1.323	**19.991**
Entertainment	7	1.176	8.232	1.320	**17.470**
Dance	6	1.477	8.862	1.344	**16.926**
drama	6	1.477	8.862	1.344	**16.926**
show	5	1.176	5.880	1.304	**12.400**
year	5	1.000	5.000	1.282	**11.409**
FNC	4	1.477	5.908	1.331	**11.232**
Mnet	4	1.477	5.908	1.331	**11.232**
style	4	1.176	4.704	1.295	**9.882**
fan	4	1.176	4.704	1.270	**9.784**
program	3	1.176	3.528	1.354	**7.589**
Seoul	3	0.875	2.625	1.351	**6.677**
scene	3	1.477	4.431	1.295	**8.315**
company	2	1.000	2.000	1.346	**4.693**
month	2	0.875	1.75	1.340	**4.429**
…	…	…	…	…	…

표 11은 TF-IDF에 편찻값을 적용한 최종 가중치 값이다. TF-IDF에서 IDF의 값이 같으므로 분별력이 없었는데, IDF에 로그 편차 값을 추가로 더함으로써 분별력 있는 가중치 값이 측정되었다. Entertainment 카테고리에서 단어 album은 TF-IDF 가중치가 가장 높다. 하지만 단어의 편차를 적용한 결과 단어 band가 해당 카테고리에서 가장 높은 가중치로 변경되었다.

4. 비교 실험

본 논문의 방법과 TF-IDF 방법으로 구축된 사전데이터의 가중치 값을 이용해 실험문서의 카테고리 분류를 한다.

표 12. 실험 문서에서 추출된 명사 예시

	추출된 명사
실험 문서1	singer, Zico, Block, AOA, eolhyun, agency, Tuesday, attention, Zico, relationship, reasons, Zico, agency, Entertainment, nature, FNC, Entertainment, Seolhyun, agency, news, attention, relationship, rapper, couple, Zico, Seolhyun, relationship, online, media, outlet, Dispatch, photo, couple,August, March

실험 문서1의 추출된 명사는 사전데이터에 있는 각 카테고리의 명사 집합과 일치여부를 파악하여, 일치된 단어의 가중치 값의 합이 높은 카테고리로 분류된다. 본 논문의 방법으로 가중치를 측정한 실험한 방법과 TF-IDF 방법으로 가중치를 실험한 결과는 표 13, 표 14와 같다.

표 13. TF-IDF를 이용한 실험 문서의 카테고리 분류 예시

	Entertainment	Traditional	Sport	Game
singer	4.804	2.425	0	0
Zico	0	0	0	0
Bolck	0	0	0	0
AOA	0	0	0	0
Seolhyun	0	0	0	0
agency	3.903	4.682	0	0
Tuesday	5.836	0	0	0
attention	0	14.434	0	1.477
relationship	0	2.382	0	0
reason	0	4.624	0	0
Entertainment	8.232	0	0	3.882
nature	0	0	0	0
FNC	5.908	0	0	0
...	0	0	0	0
합계	28.317	**28.547**	0	5.359

표 14. 본 논문 방법을 이용한 실험 문서의 카테고리 분류 예시

	Entertainment	Traditional	Sport	Game
singer	12.173	4.830	0	0
Zico	0	0	0	0
Bolck	0	0	0	0
AOA	0	0	0	0
Seolhyun	0	0	0	0
agency	7.425	7.502	0	0
Tuesday	14.788	0	0	0
attention	0	3.0152	0	2.942
relationship	0	4.744	0	0
reason	0	9.211	0	0
Entertainment	18.185	0	0	8.109
nature	0	0	0	0
FNC	14.970	0	0	0
...	0	0	0	0
합계	**68.143**	46.441	0	11.051

TF-IDF로 카테고리 분류를 할 경우 Traditional 로 분류 되었지만, 단어의 분포 차를 고려하여 카테고리의 주요 단어에 추가 가중치를 부여함으로써, Entertainment 카테고리로 올바르게 분류되었다.

본 논문의 방법과 TF-IDF 방법으로 Entertainment, Traditional, Sport, Game 분야의 사전데이터를 각각 구축하고, 전체 실험문서를 카테고리 분류한 결과는 표 15, 표 16과 같다.

표 15 TF-IDF 방법을 적용한 문서 분류 결과

	Entertainment	Traditional	Sport	Game
Entertainment	**476**	8	2	14
Traditional	11	**481**	0	8
Sport	4	1	**488**	7
Game	18	9	6	**467**

표 16 본 논문의 방법을 적용한 문서 분류 결과

	Entertainment	Traditional	Sport	Game
Entertainment	**486**	5	2	7
Traditional	9	**483**	0	8
Sport	4	1	**491**	4
Game	11	8	6	**475**

문서 분류결과 각 카테고리의 분별력 있는 단어에는 높은 가중치 값을 부여하고, 분별력이 없는 단어에는 낮은 가중치 값을 부여함으로써 해당 카테고리와 관련된 주요 단어들이 큰 값을 얻었다. 실험결과 TF-IDF는 정확도 95.6%, 본 논문의 방법은 정확도 96.75%로 정확도가 1.15% 향상되었다.

Ⅴ. 결론 및 향후 연구

스마트 기기의 발달로 소셜 네트워크 서비스 이용자가 증가하였고, 이로 인해 빅 데이터가 축적되었다. 이러한 빅 데이터를 효율적으로 처리, 분석하기 위한 방법으로 패턴 인식, 기계 학습, 자연어 처리 등이 있다. 이 중 자연어 처리 분야는 인간의 언어를 컴퓨터를 사용하여 처리하며, 주로 사용자의 편의성에 중점을 둔 연구 분야로써 텍스트 문서를 분석하여 키워드를 추출하고 정보 검색, 문서 요약, 문서 분류 등의 연구를 하고 있다.

본 논문에서는 키워드 추출 시 문서의 분별력이 없는 단어에는 낮은 가중치를 부여하고, 문서의 특징을 잘 나타내는 단어에는 높은 가중치를 부여하여 키워드를 추출한다. 먼저 위키피디아의 redirect 기능을 이용해 신조어나, 줄임말, 동의어 등 의미가 같지만, 형태가 다른 단어를 하나의 대표단어로 추출한다. 이후, 모든 문서를 고려한 단어들의 편차를 구한다. 임계치 이상의 단어에 추가 가중치를 부여함으로써 주요 키워드에 높은 가중치 값을 부여한다. 비교실험을 위해 추출된 키워드는 본 연구의 방법과 TF-IDF 방법으로 문서의 유사도를 측정하여 분류하는 비교실험을 한다.

각 카테고리의 주요 키워드에 높은 가중치를 그렇지 않은 키워드에는 낮은 가중치를 부여함으로써, 본 논문의 방법을 적용한 실험결과가 TF-IDF의 문서분류 실험결과보다 1.15% 향상된 결과를 얻었다.

향후 연구로는 위키피디아에 나와 있지 않은 신조어나 줄임말 등을 관련 있는 다른 키워드로 확장하는 방법에 대해 연구가 필요하며, 문서의 특징을 나타내는 키워드를 잘 추출하게 된다면 문서 분류뿐만 아니라 다른 자연어 처리 분야에도 유용하게 적용할 수 있을 것이다.

【참고문헌】

[1] Snijders, Chris., Matzat, Uwe., Reips, Ulf-Dietrich., 「"Big Data": Big Gaps of Knowledge in the Field of Internet Science」, 『International Journal of Internet Science」, Vol. 7, pp. 1-5, 2012.

[2] http://www.bloter.net/archives/258481

[3] 안창원, 황승구, 「빅 데이터 기술과 주요 이슈」, 『정보과학회지』 제30권, 제6호, pp. 10-17, 한국정보과학회, 2012.

[4] 김정숙, 「빅 데이터 활용과 관련기술 고찰」, 『한국콘텐츠학회지』 제10권, 제1호, pp. 34-40, 한국콘텐츠학회, 2012.

[5] 강만모, 김상락, 박상무, 「빅 데이터의 분석과 활용」, 『정보과학회지』 제30권, 제6호, pp. 25-32, 한국정보과학회, 2012.

[6] 박경미, 황규백 「자연어처리 기반 바이오 텍스트 마이닝 시스템」, 『정보과학회논문지』 제17권, 제4호, pp. 205-213, 한국정보과학회, 2011.

[7] https://ko.wikipedia.org/wiki/자연언어처리

[8] Jeremy, Ginsberg., Matthew H, Mohebbi., Rajan S, Patel., Lynnette Brammer., Mark S, Smolinski., Larry Brilliant., 「Detecting influenza epidemics using search engine query data」, 『Nature Letters』 Vol. 457, pp. 1012-1014, Nature, 2009.

[9] https://en.wikipedia.org/wiki/Keyword_extraction

[10] https://en.wikipedia.org/wiki/Text_mining

[11] https://en.wikipedia.org/wiki/Data_mining

[12] https://en.wikipedia.org/wiki/Machine_learning

[13] https://ko.wikipedia.org/wiki/정보검색

[14] R, Robertson., 「Understanding inverse document frequency: on theoretical arguments for IDF」, 『Journal of Documentation』, Vol. 60, No. 5, pp. 503-520, 2004.

[15] 이용훈, 이상범, 「Okapi BM25 단어 가중치법 적용을 통한 문서 범주화의 성능 향상」, 『한국산학기술학회논문지』 제11권, 제12호, pp. 5089-5096, 한국산학기술학회, 2010.

[16] 유은순, 최건희, 김승훈,「TF-IDF와 소설 텍스트의 구조를 이용한 주제어 추출 연구」,『한국컴퓨터정보학회논문지』제20권, 제2호, pp. 121-129, 한국컴퓨터정보학회, 2015.

[17] 박호식,「의료 정보 추출을 위한 TF-IDF 기반의 연관규칙 분석 시스템」, 아주대학교 일반대학원: 컴퓨터공학과, 석사학위논문, 2016.

[18] 이재욱, 고병규, 김판구,「상호 정보량과 로그 정규화를 이용한 뉴스 카테고리 분류」,『한국정보기술학회논문지』제14권, 제7호, pp. 79-85, 한국정보기술학회, 2016.

[19] Larocca, Neto., Joel.,「A Text Mining Tool for Document Clustering and Text Summarization」,『Proceedings of The Text Mining Tool for Document Clustering and Text Summarization Fourth International Conference on The Practical Application of Knowledge Discovery and Data Mining』, pp. 41-56, 1997.

[20] https://ko.wikipedia.org/wiki/코사인 유사도

[21] Yonghoon Lee,「A Study on Document Categorization Methods using word weighting technique」,『Department of Computer Science Graduate School Dankkook University』, 2010.

[22] Yang, Y., J, O, Pederson.,「A comparative study on feature selection in text categorization」,『Proceedings of the 14th International Conference on Machine Learning』, 1997.

[23] http://terms.naver.com/entry.nhn?docId=511803&cid=42126&categoryId=42126

[24] 김진상, 최상열,「카이제곱 통계량을 이용한 개선된 베이지안 스팸메일 필터」,『한국지능시스템학회 할술발표 논문집』제15권, 제1호, pp. 403-414, 한국지능시스템학회, 2005.

[25] 이재윤,「상호정보량의 정규화에 대한 연구」,『한국문헌정보학회지』제37권, 제4호, pp. 177-198, 한국문헌정보학회, 2003.

[26] 허정, 장명길,「평균 상호정보량에 기반한 동음이의어 중의성 해소」,『한국정보과학회 언어공학연구회 학술발표 논문집』제2005권, 제10호, pp. 159-166, 한국정보과학회 언어공학연구회, 2005.

[27] http://www.d.umn.edu/~tpederse/Group01/WordNet/words.txt

[28] W.Bruce Croft, Donald Metzler, Tervor Strohman, 『Search Engines』, PEARSON, 2010.

[29] http://www.shineware.co.kr

[30] http://www.ibm.com/support/knowledgecenter/ko/SS5RWK_3.5.0/com.ibm.discovery.es.ta.doc/iiysspostagset.htm

[31] https://en.wikipedia.org/wiki/WordNet

[32] https://en.wikipedia.org/Wikipedia

[33] http://www.koreatimes.co.kr