The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.2.265
32(2), 265–276

Comparison of term weighting schemes for document classification

Ho Young Jeong[a] · Sang Min Shin[b] · Yong-Seok Choi[a,1]

[a]Department of Statistics, Pusan National University;
[b]Department of Management Information Systems, Dong-A University

(Received January 3, 2019; Revised February 8, 2019; Accepted February 12, 2019)

Abstract

The document-term frequency matrix is a general data of objects in text mining. In this study, we introduce a traditional term weighting scheme TF-IDF (term frequency-inverse document frequency) which is applied in the document-term frequency matrix and used for text classifications. In addition, we introduce and compare TF-IDF-ICSDF and TF-IGM schemes which are well known recently. This study also provides a method to extract keyword enhancing the quality of text classifications. Based on the keywords extracted, we applied support vector machine for the text classification. In this study, to compare the performance term weighting schemes, we used some performance metrics such as precision, recall, and F1-score. Therefore, we know that TF-IGM scheme provided high performance metrics and was optimal for text classification.

Keywords: term weighting, document classification, text mining, TF-IDF, keyword extraction

1. 서론

최근에는 신문이나 블로그, 이메일 등으로부터 텍스트나 이미지의 비정형 데이터가 방대하게 생산되고 있다. 텍스트 마이닝은 자연어 처리 기술을 활용하여 비정형의 텍스트 데이터를 정형화하고 유용한 가치와 의미있는 정보를 획득할 수 있도록 하는 데이터 마이닝 기법이라 할 수 있다. 텍스트 마이닝에서의 일반적인 자료는 각각의 문서별로 특정 용어의 사용 빈도를 나타내는 문서-용어 빈도행렬(document-term frequency matrix)로 표현되는데, 이러한 문서-용어 빈도행렬의 각각의 행은 개별 문서를 의미하고 각각의 열은 특정 용어들을 의미한다.

하지만 문서-용어 빈도행렬에서 표현되는 용어들의 빈도만을 가지고 문서들의 차별성과 용어들의 중요도를 나타내기 어려우므로, 문서의 특징을 분류하기 위해서는 용어 가중치(term weighting)를 이용하는 것이 필수적이다. 용어 가중치는 정보 검색이나 텍스트 마이닝 분야에 이용되고 있으며 여러 문서로 이루어진 문서집단이 있을 때 어떤 용어가 특정 문서 내에서 얼마나 중요한지를 나타내는 통계적 수치이다. 따라서 효과적인 용어 가중치를 이용한다면 비정형 자료로부터 더욱 유용한 가치와 의미를 도출함과 동시에 문서 분류에 있어서 더 좋은 결과를 얻을 수 있다.

[1]Corresponding author: Department of Statistics, Pusan National University, 2, Busandaehak-ro 63beon-gil, Geumjeong-Gu, Busan 46241, Korea. E-mail: yschoi@pusan.ac.kr

용어 가중치에 대한 연구는 현재까지 다양한 분야에서 연구되어 왔다. 가장 대표적인 용어 가중치인 term frequency-inverse document frequency (TF-IDF)는 컴퓨터공학, 정보통신공학, 문헌정보학뿐만 아니라 생물학, 의학, 인문학 등 모든 분야에서 다양하게 활용할 수 있다. 텍스트 마이닝과 관련된 지금까지의 선행연구들은 대다수 이러한 TF-IDF와 같은 용어 가중치를 이용하여 다수의 문서들에 대한 특징을 파악하고 주어진 문서들을 분류하는 것을 목적으로 하고 있다. 그러나 다수의 개체(object)들이 각각 둘 이상의 문서를 발행한 경우에는 발행된 문서들에 대한 특징을 파악하고 주어진 문서들을 분류하는 것에만 목적을 두는 것이 아니라, 문서들의 특징을 이용하여 개체들에 대한 분류에도 목적을 둘 필요가 있다. 따라서 본 연구에서는 다수의 개체들이 각각 둘 이상의 문서를 발행한 경우에 문서 및 개체의 특징을 파악하기 위한 다양한 용어 가중치들을 소개하고 이들의 계산법과 장단점을 정리하여 간단한 예시와 함께 이해를 돕고자한다.

이를 위해 2장에서는 기존의 문서-용어 빈도행렬에 개체 정보를 추가한 새로운 문서-용어 빈도행렬을 정의하고, 이에 대해 적용가능한 다양한 용어 가중치들의 특징을 소개하고자 한다. 그리고 3장에서는 활용사례를 이용하여 비정형의 텍스트 데이터를 정형화하는 과정과 정형화된 텍스트 데이터에 대해 각각의 용어 가중치를 적용하는 과정을 기술하며, 문서 분류에서 가장 많이 이용되는 서포트 벡터 머신(support vector machine; SVM)을 적용하여 용어 가중치들 중에서 문서 및 개체 분류에 대해 최적화된 방법을 찾아보고자 한다. 끝으로 4장의 결론에서 본 연구를 정리 및 요약한다.

2. 문서 분류를 위한 용어 가중치

2.1. 텍스트 자료의 정형화

특정 문서나 웹 페이지에서 원하는 텍스트 데이터를 추출하는 행위를 크롤링(crawling)이라고 한다. 그리고 크롤링을 이용하여 텍스트 데이터를 추출하게 되면, 대용량의 텍스트 집합이 생성되는데 이를 Miner 등 (2012)은 말뭉치(corpus)로 정의하였다. 말뭉치는 비정형 자료이기 때문에 정형화된 자료로 변환시켜 주어야 하는데, 이를 위해 우선 문장부호, 특수문자, 불용어(stop words) 등의 제거와 같은 정제(cleaning) 과정이 필요하다. 다음으로 정제 과정을 거친 텍스트 데이터를 정형화하는 과정에서는 가장 대표적인 방법으로 벡터 공간 모델(vector space model)을 이용하는데 벡터 공간 모델에서 문서는 단어 주머니(bag of words) 역할을 하는 벡터로 표현되며, 이들 벡터의 차원은 개별 용어에 대응된다. 따라서 각각의 벡터들은 대응되는 문서에서 특정 용어의 출현 빈도를 원소값으로 가지게 되며, 이러한 벡터들을 이용하여 정형화된 문서-용어 빈도행렬을 생성할 수 있다.

만약, g개의 개체들이 각각 n_r개의 문서를 가지고 있다면, 전체 문서수는 $n = \sum_{r=1}^{g} n_r$이 된다. 따라서 n개 문서에 대한 p개 용어로 이루어진 크기 $n \times p$의 문서-용어 빈도 행렬은 다음과 같이 정의할 수 있다.

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_g \end{bmatrix} = (y_{ij}^r), \quad i = 1, \ldots, n_r;\ j = 1, \ldots, p;\ r = 1, \ldots, g, \tag{2.1}$$

여기서 y_{ij}^r은 r번째 개체의 i번째 문서에서 j번째 용어의 출현 빈도를 나타내며, $\mathbf{Y_r}$은 크기 $n_r \times p$의 행렬로 행렬 $\mathbf{Y}$의 r번째 부분행렬이다. 따라서 $\mathbf{Y_r}$는 r번째 개체에 대한 문서-용어 빈도행렬이다.

2.2. 용어 가중치 함수

단순히 어느 문서에서나 등장할 법한 일반적인 용어(예를 들어 '일반', '방법', '영향' 등과 같은 일반 명

사 용어)의 빈도가 높다고 해서 개체를 대표하는 용어라고 판단하기 어렵다. 따라서 식 (2.1)에서 정의한 문서-용어 빈도행렬 $\mathbf{Y}$은 용어의 중요도를 반영하지 못하기 때문에 이를 수치화할 필요가 있다. 이때, 문서와 용어의 상호 관계를 설명하기 위해 용어 가중치 함수를 이용하게 되는데 용어 가중치 함수는 지역적 가중치 함수(local weight function)와 전역적 가중치 함수(global weight function)로 구분된다.

지역적 가중치 함수는 i번째 문서에서 추출한 j번째 용어의 발생빈도에 가중치를 부여하는 함수로, Nakov 등 (2001)은 전체 용어의 수가 많지 않은 상황에서 i번째 문서에서 추출한 j번째 용어가 한 번이라도 나타나는 경우는 1, 나타나지 않는 경우는 0으로 할당하는 불린(boolean) 방법과 용어의 발생빈도에 1을 더한 뒤에 로그 값을 취한 로그 변환(logarithm) 방법을 사용하였다. 특히, Chen과 Zong (2003)은 용어의 발생빈도에 대한 제곱근을 이용하면 로그 변환 방법보다 문서 분류에 있어서 동등하거나 좋은 결과를 가져옴을 보인 바 있다.

그러나 특정 용어가 모든 문서에서 같이 나타난다면 지역적 가중치 함수만으로는 각각의 문서나 개체를 구별하지 못하는 한계점이 있다. 이러한 경우 전역적 가중치 함수를 이용하면 문서와 개체 전체에 대한 용어의 특징을 반영할 수 있다. 전역적 가중치 함수는 문서 또는 개체 전체에 대한 j번째 용어에 가중치를 부여하는 함수로, Dumais (1991)는 용어 빈도에 대해 표준화를 시킨 정규화(normal) 방법과 문서 전체에서 용어의 빈도에 대한 용어가 나타난 문서 수의 비율을 고려한 global frequency inverse document frequency (Gfidf) 방법 등을 사용하였다.

이 절에서는 용어의 정보를 양적으로 나타낸 대표적인 용어 가중치인 TF-IDF와 최근에 제안된 용어 가중치인 TF-IDF-ICSDF, TF-IGM을 소개하고자 한다. 이를 위해 우선, r번째 개체의 i번째 문서에서 추출한 j번째 용어의 출현빈도에 대한 지역적 가중치를 $L_r(i,j)$라 하고, 문서 또는 개체에 대한 j번째 용어의 전역적 가중치를 $G(j)$라고 하면, r번째 개체의 i번째 문서에서 추출한 j번째 용어에 대한 문서-용어 가중점수를 식 (2.2)와 같이 정의하자.

$$z_{ij}^r = L_r(i,j) \times G(j) \tag{2.2}$$

그리고 문서-용어 빈도행렬 $\mathbf{Y}$에 대해 용어 가중치를 적용한 크기 $n \times p$의 문서 용어 가중행렬을 $\mathbf{Z} = (z_{ij}^r),\ i = 1, \ldots, n_r;\ j = 1, \ldots, p;\ r = 1, \ldots, g$로 정의한다.

2.2.1. TF-IDF TF-IDF는 특정 용어의 중요도는 용어가 출현한 횟수에 비례하고 그 용어가 언급된 모든 문서의 총 수에 반비례한다는 명제에 기초하고 있다 (Lee와 Bae, 2002). TF-IDF에 의한 가중점수 z_{ij}^r를 산출하기 위한 지역적·전역적 가중치 함수는 식 (2.3)과 같이 정의된다.

$$L_r(i,j) = y_{ij}^r, \qquad G(j) = \log\left(\frac{n}{DF(n,j)}\right), \tag{2.3}$$

여기서 y_{ij}^r는 i번째 문서에서 추출한 j번째 용어의 발생빈도(term frequency)이며, $G(j)$항을 IDF라고 부른다. DF(n,j)는 n개의 문서 중 j번째 용어가 포함된 문서의 수(document frequency)를 나타내고 n은 전체 문서의 수이다. 식 (2.3) 로그 값의 분모인 DF(n,j)가 0일 경우 n개의 문서에서 j번째 용어가 한 번도 등장하지 않은 상황이므로 y_{ij}^r값 또한 0이 되기 때문에 값은 0으로 간주한다. DF$(n,j)/n$은 전체 n개의 문서에 대한 j번째 용어가 포함된 문서의 수의 비율이다. 이 비율이 높을수록 여러 문서에 용어들이 많이 등장한다는 의미가 되고, 용어의 중요도는 낮아지므로 역수 변환을 하고 이 영향을 줄이기 위해 로그 변환을 고려한다.

TF-IDF는 특정한 문서에서 많이 등장하는 용어일수록 해당문서의 특성이 되는 용어이므로 높은 가중치를 얻는다. 그러나 모든 문서에서 등장하는 중요도가 낮은 용어는 $\log 1$의 값을 가져 가중치가 0이 된

다. 예를 들어 어느 문서에서나 등장할 법한 일반적인 용어의 가중치는 상대적으로 작을 것이고 가중치가 0이 되면 용어 집합에서 제외할 수 있음을 의미한다.

2.2.2. TF-IDF-ICSDF IDF의 단점은 개별 문서에 대해서만 용어의 대표성을 반영한다. 즉 문서와 용어 사이의 정보만을 표현할 뿐 개체들의 정보를 무시하는 문제점이 있다. 이를 보완하고자 Ren과 Sohrab (2013)이 소개한 TF-IDF-inverse class space density frequency (TF-IDF-ICSDF) 용어 가중치에 대해 설명을 하고자 한다. TF-IDF-ICSDF에 의한 가중점수 z_{ij}^r를 산출하기 위한 지역적·전역적 가중치 함수는 식 (2.4)와 같이 정의된다.

$$L_r(i,j) = y_{ij}^r, \qquad G(j) = \log\left(\frac{n}{\mathrm{DF}(n,j)}\right) \times \log\left(\frac{g}{\mathrm{CS}(j)}\right). \tag{2.4}$$

제 2.2.1절에서의 TF-IDF 용어 가중치에 개체들의 정보를 부여하기 위해 추가적으로 $\log(g/\mathrm{CS}(j))$를 곱해주면 되고 이 항을 ICSDF 라고 부른다. 용어들과 개체에 대한 정보를 반영하기 위해 각 개체에서 해당용어가 나타내는 확률의 합계를 다음과 같이 정의한다.

$$\mathrm{CS}(j) = \sum_{r=1}^{g} \frac{d_{rj}}{n_r}. \tag{2.5}$$

식 (2.5)에서 g는 문서-용어 빈도행렬의 총 개체의 수이다. d_{rj}은 j번째 용어가 적어도 한 번이라도 나타난 r번째 개체에 들어 있는 문서의 수이고, n_r은 r번째 개체 내의 총 문서의 수를 의미한다. $\mathrm{CS}(j)$는 class space density의 약어이며 각 개체에서 해당 용어가 나타나는 확률의 합계로 계산된다. TF-IDF-ICSDF는 하나의 개체를 대표하는 용어는 다른 개체에서 가끔씩 출현하는 것보다 더 많은 가중치가 부여하기 때문에 개체의 대표성을 잘 나타낼 수 있는 가중치이다. 하지만 상대적으로 대표성이 없고 여러 번 나타나는 용어들에 대해서 $\mathrm{CS}(j)$값이 커지므로 낮은 가중점수를 가지게 되어 개체들을 대표하는 용어들을 구별하는 데 있어서 나쁜 결과를 가지게 된다.

2.2.3. TF-IGM Chen 등 (2016)은 TF-IDF-ICSDF 용어 가중치는 개체 내(within a class)의 관점으로 계산하는 가중치이므로 개체 내에서 문서의 수가 균등하게 나타나는 경우나 개체 내의 총 문서의 수의 차이가 클 때 개체들을 구별하는 능력이 부족하여 좋은 결과를 얻기 어렵다고 주장했다. 개체 간(across different classes)의 관점으로 계산한 용어 가중치 Chen 등 (2016)이 소개한 TF-inverse gravity moment (TF-IGM)는 개체 구분을 보다 정확하게 계산하는 용어 가중치이다. TF-IGM에 의한 가중점수 z_{ij}^r를 산출하기 위한 지역적·전역적 가중치 함수는 식 (2.6)과 같이 정의된다.

$$L_r(i,j) = y_{ij}^r, \qquad G(j) = \mathrm{IGM}(j). \tag{2.6}$$

문서-용어 빈도행렬 $\mathbf{Y}$에서 용어에 대한 가중점수를 계산하기 위해 원소 y_{ij}^r에 가중치 $G(j)$를 부여하여 가중점수 z_{ij}^r를 산출한다. 가중치 IGM을 다음과 같이 정의한다.

$$\mathrm{IGM}(j) = 1 + \lambda \times \frac{d_{(1)j}}{\sum_{r=1}^{g} d_{rj} \times R_{rj}}. \tag{2.7}$$

식 (2.7)에서 j번째 용어가 적어도 한 번이라도 나타난 r번째 개체에 들어 있는 문서의 수는 d_{rj}이므로 개체에 있는 총 문서들의 수는 $d_{1j}, d_{2j}, \ldots, d_{gj}$이다. 이를 내림차순으로 정렬하면 $d_{(1)j} \geq d_{(2)j} \geq \cdots \geq d_{(g)j}$로 나타낼 수 있고, $d_{(1)j}$는 j번째 용어가 포함된 개체 중에 가장 많이 출현한 문서의 수를 의미한다. R_{rj}는 $d_{1j}, d_{2j}, \ldots, d_{gj}$들의 순위를 의미하는 것이고, 동점인 경우는 평균 순위를 적용한다.

Table 2.1. IGM calculation example

Object	Term 1	Rank	Term 2	Rank
Obj1	■■■■□□□□□□	1	□□□□□□□□□□	4
Obj2	■■□□□□□□□□	3.5	■■■■□□□□□□	2
Obj3	■■□□□□□□□□	3.5	■■■■■■■■□□	1
Obj4	■■□□□□□□□□	3.5	□□□□□□□□□□	4
Obj5	■■□□□□□□□□	3.5	□□□□□□□□□□	4

Table 2.2. Document-term weighted matrix generation scheme M1–M6

Local weight function	Global weight function	Term weighting scheme	Weighted matrix
y_{ij}^r	$\log\left(\frac{n}{DF(n,j)}\right)$	[Scheme M1] TF-IDF	$\mathbf{Z_1}$
	$\log\left(\frac{n}{DF(n,j)}\right) \times \log\left(\frac{g}{CS(j)}\right)$	[Scheme M2] TF-IDF-ICSDF	$\mathbf{Z_2}$
	IGM(j)	[Scheme M3] TF-IGM	$\mathbf{Z_3}$
$\sqrt{y_{ij}^r}$	$\log\left(\frac{n}{DF(n,j)}\right)$	[Scheme M4] RTF-IDF	$\mathbf{Z_4}$
	$\log\left(\frac{n}{DF(n,j)}\right) \times \log\left(\frac{g}{CS(j)}\right)$	[Scheme M5] RTF-IDF-ICSDF	$\mathbf{Z_5}$
	IGM(j)	[Scheme M6] RTF-IGM	$\mathbf{Z_6}$

λ는 조정 가능 계수(adjustable coefficient)로써 5.0에서 9.0 사이의 값을 가진다. 본 연구에서는 조정 가능계수를 7.0으로 설정했을 때 가장 좋은 결과가 나타났다.

예를 들어 위의 Table 2.1과 같이 문서-용어 빈도행렬에 5개의 개체에 각각 동일하게 10개의 문서가 있고 개체에 특정 용어 Term t_1와 Term t_2가 나타난 문서의 수가 각각 $\{4,2,2,2,2\}$, $\{0,4,8,0,0\}$ 처럼 나타났다고 가정을 하자. 따라서 각 용어가 나타난 문서의 순위는 각각 $\{1,3.5,3.5,3.5,3.5\}$, $\{4,2,1,4,4\}$이 된다. 직관적으로 Term t_2가 세번째 개체를 대표하는 용어임을 알 수 있고 개체들을 더 강하게 구별을 할 수 있다. 식 (2.7)에서 조정 가능 계수 λ를 7로 설정하여 계산하면 Term t_1에 대한 IGM 값은 $1+7\times4/32 = 1.875$, Term t_2에 대한 IGM 값은 $1+7\times8/16 = 4.5$로 나타났기 때문에 Term t_2 에 대한 IGM 값이 높게 나옴을 알 수 있다. 이를 통해 IGM값은 개체를 대표하는 용어에 더 높은 가중점수를 주게 되어 개체를 더 강하게 구별을 한다는 것을 알 수 있다. 같은 조건에서 앞서 언급했던 전역적 가중치 함수인 IDF, ICSDF 값은 Term t_1와 Term t_2에 대해서 모두 1.427로 똑같은 가중점수를 주기 때문에 개체들을 구별하는데 합리적이지 못함을 알 수 있다. 하지만 IGM 값은 두 용어에 대해 다른 가중점수를 주기 때문에 개체를 구분하는 데 있어서 합리적임을 알 수 있다.

Chen과 Zong (2003)은 용어의 발생빈도에 대한 제곱근을 이용하면 Nakov 등 (2001)의 로그 스케일 방법보다 문서 분류에 있어서 동등하거나 좋은 결과를 가져옴을 보인 바 있다. 위에서 언급한 모든 가중치의 $L_r(i,j)$는 i번째 문서에서 추출한 j번째 용어의 발생빈도인 y_{ij}^r를 그대로 사용하였다. 본 연구에서 y_{ij}^r값의 영향력을 줄이기 위해 $L_r(i,j)$를 제곱근을 씌운 $\sqrt{y_{ij}^r}$에 이번 절에 설명한 전역적 가중치 함수들을 곱으로 조합한다. 이렇게 조합한 가중치들을 RTF-IDF, RTF-IDF-ICSDF, RTF-IGM으로 새롭게 정의한다. 따라서 본 연구에 사용할 모든 용어 가중치는 Table 2.2와 같이 정리할 수 있다. 문서-용어 빈도행렬 에 M1–M6을 모두 적용한 문서-용어 가중행렬을 $\mathbf{Z_1}$–$\mathbf{Z_6}$로 정의한다.

2.3. 문서-핵심어 가중행렬

2.2절에서 생성된 문서-용어 가중행렬들은 문서-용어 빈도행렬에서의 모든 용어를 포함하고 있기 때문에 차원 수가 크고 0의 값이 많은 희소(sparse) 행렬이다. 일반적으로 문서의 수보다 용어의 수가 매우

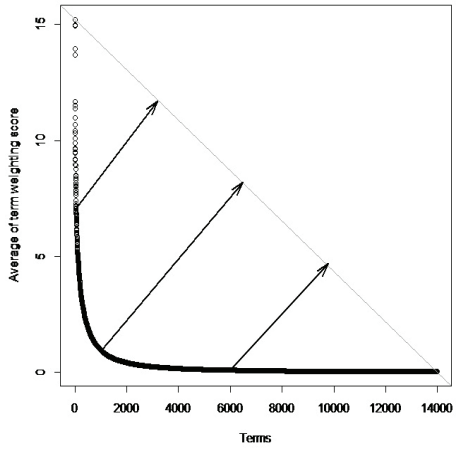

Figure 2.1. The process of finding an elbow point.

크기 때문에 문서-용어 가중행렬을 그대로 분석을 하게 된다면 많은 시간이 소요되고 분석의 질을 저하시키는 일이 생긴다. 용어의 수를 줄여야 하는 이유는 모델 해석을 단순화시켜주고, 차원 수를 줄일 경우 모델의 정확도가 증가할 수 있기 때문이다.

핵심어를 선정하는 작업을 텍스트 마이닝에서 의미 정보 추출(feature selection) 또는 용어 필터링(filtering)이라고 한다. Cho 등 (2015)은 문서-용어 가중행렬 $\mathbf{Z}$에서 각 용어들이 갖는 평균 가중점수를 계산한 후 점수가 급격하게 감소하는 지점, 팔꿈치 지점(elbow point)을 기준으로 차원을 축소하는 방법을 제안하였다. 팔꿈치 지점이 명확하지 않을 때에는 해당 분야 전문지식이 있는 사람들의 주관적인 판단에 맡기므로 팔꿈치 지점의 선택은 누가 분석하느냐에 따라서 선정되는 핵심어가 달라진다. Satopaa 등 (2011)은 Kneedle algorithm을 활용하여 그래프의 변곡점을 찾고 이 지점을 팔꿈치 지점으로 정의한다. 따라서 평균 가중점수가 급격하게 변하는 팔꿈치 지점으로 간주할 수 있고 핵심어를 선정하는 방법에 대한 객관성을 확보할 수 있다. 핵심어 선정을 위한 용어 필터링 방법은 먼저 문서-용어 가중행렬 에서 각 용어 평균 가중점수로 이루어진 평균 벡터 $\bar{\mathbf{z}}$를 계산한다.

$$\bar{\mathbf{z}} = (\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_p)^t = \frac{1}{n}\mathbf{Z}^t\mathbf{1}_n. \tag{2.8}$$

$\bar{\mathbf{z}}$의 원소를 큰 값에서 작은 값 순으로 내림차순하면 $\bar{z}_{(1)} \geq \bar{z}_{(2)} \geq \ldots \geq \bar{z}_{(p)}$와 같으며 이를 바탕으로 그래프를 그린 후 그래프의 처음과 끝 지점인 $\bar{z}_{(1)}$과 $\bar{z}_{(p)}$를 관통하는 직선을 이어주면 Figure 2.1과 같다.

Figure 2.1에서 화살표로 표시한 것과 같이 그래프에서 직선까지 이르는 거리가 최대가 되는 그래프 상의 지점을 변곡점, 즉 팔꿈치 지점으로 간주할 수 있다. 팔꿈치 지점을 기준으로 평균 가중점수가 높은 상위 용어 개를 핵심어로 선정한다. 이를 통해 행렬 $\mathbf{Z}$로부터 문서-핵심어 가중행렬 $\mathbf{X} = (x_{rlt})$, $t = 1, \ldots, q$를 얻게 된다. 따라서 2.2절에서 언급한 Table 2.2의 문서-용어 가중행렬 $\mathbf{Z_1}$–$\mathbf{Z_6}$에 용어 필터링 방법을 적용한 문서-핵심어 가중행렬을 $\mathbf{X_1}$–$\mathbf{X_6}$로 정의한다.

3. 활용 사례

3.1. 자료 수집 및 생성

본 연구의 자료 수집은 Jung (2017)의 연구 자료를 인용하며 일부분을 사용하였다. 본 연구에서는 분석

Table 3.1. PDF files and terms of Periodical publication by institute

	Korea institute	Periodical publication	Number of File	Number of Terms
1	대외경제정책연구원(A)	정책연구브리핑	23	4,252
2	한국보건사회연구원(B)	보건복지포럼	132	9,899
3	한국청소년정책연구원(C)	한국청소년연구	39	6,262
4	환경정책평가연구원(D)	환경정책	30	6,790
Sum			224	27,203
Number of terms after deduplication				14,474
Number of terms after delete stopwords				13,980

대상이 되는 정부출연연구기관을 경제·인문사회연구회 소속 26개 연구기관으로 한정함으로써 한국의 경제 및 사회 분야로 분석 범위를 좁히고자 한다. 2016년 동안 발간된 정기간행물 중에서 기관별 대표 정기간행물을 다음과 같은 기준으로 선별하였다.

1. 각 기관의 홈페이지에 PDF 파일의 형식으로 공개되어 저장이 가능한 간행물
2. PDF 파일이 암호화되지 않아 한국어 텍스트 추출이 가능한 국문 간행물
3. 각 기관의 연구결과를 종합적으로 제시하는 간행물
4. 기관에서 발행하는 간행물 파일의 수가 20개 이상인 간행물

선별기준에 적합한 간행물 파일을 제공하지 않는 7개 기관과 정기 간행물의 문서의 수가 적다고 판단한 간행물 15개 기관을 제외한 4개 연구기관을 분석대상으로 선정하였다. 각 기관에서의 간행물 문서 수가 적은 간행물을 제외한 이유는 3.3절에서 훈련과 테스트를 하는 데 있어서 샘플이 다소 부족하다고 판단하여 제외하였다.

분석의 대상이 되는 개체는 대외경제정책연구원, 한국보건사회연구원, 한국청소년정책연구원, 한국정책평가연구원 4개의 연구기관이다. 대외경제정책연구원에서 발행되는 간행물은 '정책연구브리핑'으로 주요연구보고서의 연구결과를 제시하거나 정책 시사점을 정리하였다. 한국보건사회연구원에서 발행되는 간행물은 '보건복지포럼'으로 보건복지부문의 정책과제분석 결과 및 정책 동향을 전달하는 역할을 한다. 한국청소년정책연구원에서 발행되는 간행물은 '한국청소년연구'로 한국연구재단에 등재된 청소년연구 관련 전문학술지이다. 마지막으로 한국환경정책평가연구원에서 발행되는 간행물은 '환경정책'으로 환경정책 및 현안에 대한 정보를 수록한 학술지이다. 개체의 연구기관에서의 정기간행물들의 발간목적과 연구분야가 모두 다름을 알 수 있다.

본 연구에서는 파이썬(Python) 프로그램을 이용하여 각 기관에서 수집된 간행물 PDF 파일로부터 텍스트를 크롤링하여 말뭉치를 TXT 파일로 저장한다. 그리고 한국어 자연어 처리(Natural Language Processing; NLP)를 위한 파이썬에서 제공하는 패키지 Korean NLP in Python (KoNLPy)를 이용하였고 형태소 분석기는 꼬꼬마(Kkma) 형태소 분석기를 사용하였다. 텍스트 자료를 정형화하기 위해 먼저 말뭉치에서 띄어쓰기, 문장부호, 특수문자, 영문 및 기호 등을 제거하는 정제화 작업을 시행한다. 그 이후 한국어 형태소 품사 중에서도 체언에 해당하는 일반 명사와 고유 명사만을 추출하여 용어로 활용하였다. 그리고 불용어 격인 '노, 동, 인' 등과 같은 한 글자 용어 494개를 삭제하였다. 최종적으로 분석에 사용한 기관별 대표 정기간행물 PDF 파일 개수와 파일에서 추출한 용어의 개수를 Table 3.1로 정리한다.

말뭉치에서 정제화 및 용어를 추출하는 과정 이후로는 R 프로그램을 이용하였고 수집한 자료를 바탕으로 2.1절의 개체 정보가 존재하는 문서-용어 빈도행렬을 생성한다. 분석의 대상이 되는 개체는 경제·인

문사회연구회 소속 4개 연구기관이며 빈도행렬의 행에 해당하는 문서는 2016년 한 해 동안 발간된 연구기관별 간행물 224개, 열에 해당하는 용어는 13,980개의 명사 용어가 된다. 이를 바탕으로 생성한 문서-용어 빈도행렬에 Table 2.1의 M1–M6 방법을 적용하여 문서-용어 가중행렬 $\mathbf{Z_1}$–$\mathbf{Z_6}$을 생성한다. 여기에 2.3절에서 설명한 용어 필터링 방법을 적용한 문서-핵심어 가중행렬 $\mathbf{X_1}$–$\mathbf{X_6}$을 바탕으로 SVM을 적용하여 최종적으로 정기간행물들이 각 연구기관의 발간목적에 맞게 제대로 분류가 되었는지 용어 가중치들의 성능을 보고자 한다.

3.2. 성능 평가 지표

문서 분류 결과를 평가하는데 가장 대표적인 성능평가 지표는 정확률(precision), 재현율(recall), F_1 점수를 이용한다. 또한 목표가 되는 범주를 긍정(positive) 범주라고 하며, 다른 범주들은 모두 부정(negative) 범주라고 부른다. 범주의 수가 g개인 범주를 $C_1, C_2, \ldots, C_g$라고 하고 목표가 되는 범주를 r번째 범주인 C_r이라고 했을 때 개별 범주에 대한 정확률, 재현율, F_1 점수는 아래와 같이 정의한다.

$$P(C_r) = \frac{\mathrm{TP}(C_r)}{(\mathrm{TP}(C_r) + \mathrm{FP}(C_r))}, \tag{3.1}$$

$$R(C_r) = \frac{\mathrm{TP}(C_r)}{(\mathrm{TP}(C_r) + \mathrm{FN}(C_r))}, \tag{3.2}$$

$$F_1(C_r) = 2\times\frac{P(C_r)\times R(C_r)}{P(C_r) + R(C_r)}. \tag{3.3}$$

정확률은 목표 범주를 예측할 때 예측이 얼마나 정확한가를 나타내고 r번째 범주에 대한 정확률은 식 (3.1)로 계산한다. 재현율은 분류 결과 중 실제 값의 범주의 비율을 나타내고 r번째 범주에 대한 재현율은 식 (3.2)로 계산한다. r번째 범주에 대한 F_1점수는 식 (3.3)과 같이 r번째 범주에 대한 정확률과 재현율의 조화평균으로 계산한다.

그리고 전체 범주의 성능을 평가하기 위해 매크로평균(macro average)과 마이크로평균(micro average)을 지표로 이용한다. Yang과 Liu (1999)은 두 지표에 대해 매크로 평로은 빈도의 수가 적은 범주에 영향을 받는 성능 평가 지표이고 마이크로평균은 빈도의 수가 많은 범주에 영향을 받는 성능 평가 지표라고 주장했다.

$$F_1^{\mathrm{macro}} = \frac{1}{g}\sum_{r=1}^{g} F_1(C_r). \tag{3.4}$$

매크로평균 F_1점수의 경우 식 (3.3)의 개별 범주에 대한 F_1점수를 더한 다음 모든 범주 수로 나누어 평균을 계산하는 방법으로 식 (3.4)와 같이 정의된다.

마이크로평균 F_1점수를 계산하기 위해선 마이크로평균 정확률과 마이크로평균 재현율이 필요하다. 먼저 모든 범주에 대해 TP, FP, FN 값을 계산한 후 개별 범주에 대한 정확률과 재현율을 계산했을 때와 똑같은 방식으로 계산할 수 있다. 마이크로평균 정확률은 $P^{\mathrm{micro}} = \sum_{r=1}^{g}\mathrm{TP}(C_r)/\sum_{r=1}^{g}(\mathrm{TP}(C_r) + \mathrm{FP}(C_r))$로 정의되고 마이크로평균 재현율은 $R^{\mathrm{micro}} = \sum_{r=1}^{g}\mathrm{TP}(C_r)/\sum_{r=1}^{g}(\mathrm{TP}(C_r) + \mathrm{FN}(C_r))$로 정의된다.

$$F_1^{\mathrm{micro}} = 2\times\frac{P^{\mathrm{micro}}\times R^{\mathrm{micro}}}{P^{\mathrm{micro}} + R^{\mathrm{micro}}} \tag{3.5}$$

마이크로평균 F_1점수는 마이크로평균 정확률과 마이크로평균 재현율의 조화평균으로 식 (3.5)와 같이 정의된다. 본 연구에서는 개별 범주에 대한 정확률, 재현율, F_1점수와 전체 범주의 성능을 평가하기 위해 매크로 F_1점수와 마이크로 F_1점수를 이용한다.

Table 3.2. Performance comparison with M1–M6 method for individual categories

		Precision		recall		F1 score	
		Mean	SD	Mean	SD	Mean	SD
M1: TF-IDF	대외경제정책연구원(A)	0.9992	0.0060	0.7083	0.0599	0.8102	0.0471
	한국보건사회연구원(B)	0.8767	0.0137	0.9230	0.0106	0.8978	0.0098
	한국청소년정책연구원(C)	0.5621	0.0210	0.8839	0.0268	0.6820	0.0167
	환경정책평가연구원(D)	0.9763	0.0692	0.1820	0.0463	0.3966	0.0689
M2: TF-IDF-ICSDF	대외경제정책연구원(A)	0.9108	0.0289	0.7982	0.0518	0.8370	0.0370
	한국보건사회연구원(B)	0.7845	0.0327	0.9769	0.0134	0.8661	0.0185
	한국청소년정책연구원(C)	0.8447	0.0664	0.5934	0.1072	0.6364	0.0730
	환경정책평가연구원(D)	0.9009	0.1011	0.1760	0.0371	0.3894	0.0701
M3: TF-IGM	대외경제정책연구원(A)	0.9651	0.0239	0.9810	0.0222	0.9706	0.0191
	한국보건사회연구원(B)	0.9119	0.0183	0.9953	0.0050	0.9509	0.0104
	한국청소년정책연구원(C)	0.9683	0.0237	0.8584	0.0479	0.9001	0.0356
	환경정책평가연구원(D)	0.9815	0.0237	0.6875	0.0530	0.7946	0.0415
M4: RTF-IDF	대외경제정책연구원(A)	0.9522	0.0194	0.9350	0.0337	0.9379	0.0241
	한국보건사회연구원(B)	0.9559	0.0070	0.9998	0.0013	0.9770	0.0038
	한국청소년정책연구원(C)	0.9758	0.0196	0.9427	0.0114	0.9562	0.0123
	환경정책평가연구원(D)	0.9630	0.0132	0.8038	0.0414	0.8672	0.0297
M5: RTF-IDF-ICSDF	대외경제정책연구원(A)	0.9303	0.0129	0.9622	0.0156	0.9417	0.0106
	한국보건사회연구원(B)	0.9078	0.0062	0.9994	0.0021	0.9508	0.0035
	한국청소년정책연구원(C)	1.0000	0.0000	0.8711	0.0097	0.9272	0.0067
	환경정책평가연구원(D)	0.9575	0.0255	0.6468	0.0315	0.7580	0.0299
M6: RTF-IGM	대외경제정책연구원(A)	1.0000	0.0000	0.9879	0.0196	0.9930	0.0110
	한국보건사회연구원(B)	0.9768	0.0062	1.0000	0.0000	0.9881	0.0032
	한국청소년정책연구원(C)	1.0000	0.0000	0.9328	0.0159	0.9629	0.0093
	환경정책평가연구원(D)	0.9916	0.0132	0.9793	0.0223	0.9842	0.0140

3.3. 분류 분석 결과

문서-핵심어 가중행렬 $\mathbf{X_1}$–$\mathbf{X_6}$에서 224개의 문서의 수에 대해 8:2의 비율로 훈련과 테스트를 적용하였다. 또한 훈련 데이터의 20%는 최적의 모수를 찾기 위해 검증 데이터 집합으로 지정하였다. 분류 결과의 타당성을 높이기 위해 교차 검증을 200번 반복하였다. 또한 매회 교차검증을 반복할 때마다 훈련 샘플들을 중복되지 않게 층화추출법(stratified sampling method)을 적용하고 각 개체의 문서에 대해 골고루 훈련과 테스트를 할 수 있게 하였다.

그리고 개체의 수가 4개이므로 SVM을 적용할 때 일대일 분류 방법을 이용하였고 커널함수는 선형커널을 이용하였다. 일반적으로 분류 성능이 뛰어나고 많은 유형의 데이터에 적용이 된다고 알려진 가우시안 RBF 커널을 사용하지만 Hornik 등 (2006)은 본 연구와 같이 문서 분류 분야에서 데이터가 희소행렬인 경우에는 선형커널을 이용하는 경우가 더 좋은 결과를 가져온다고 한다. Table 2.1의 M1–M6 방법에 대하여 비용 모수를 2^4로 고정하여 SVM에 적용한 후 3.2절에서 설명한 성능 평가 지표들을 계산한다. 먼저 개별 범주 분류의 성능을 살펴보기 위해 정확률, 재현율, F_1점수 매 시행마다 계산하여 평균과 표준편차를 계산한 결과를 Table 3.2에 나타낸다.

Table 3.2에서 개별 범주에 대한 지표들을 살펴보면 M1 방법에서 C개체에 대하여 0.5621의 낮은 정확률을 보이고 M1, M2 방법에서 D개체에 대해 매우 낮은 재현율(0.1820, 0.1760)을 보인다. 지역적 가중치함수를 빈도를 이용한 M1–M3 방법에서 특정 개체에 대해 높은 표준편차 값을 가지기 때문에 상대

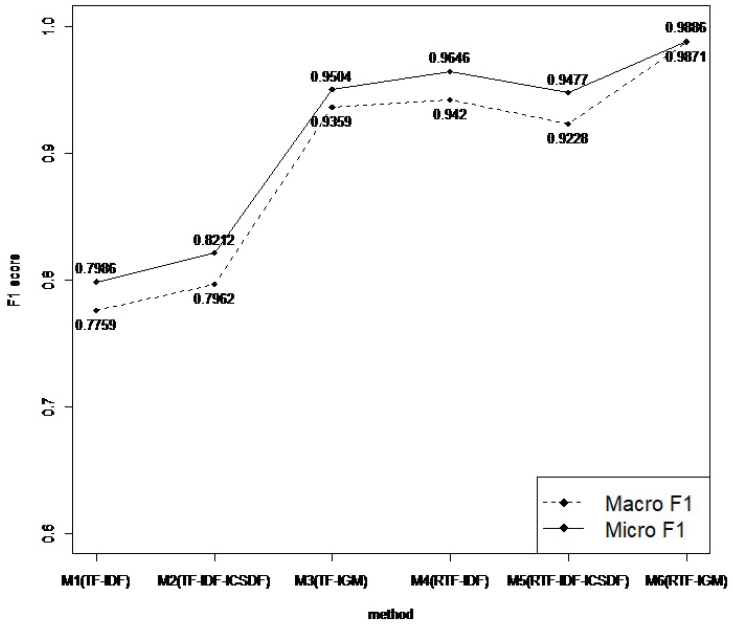

Figure 3.1. Performance comparison with M1–M6 method for entire categories.

적으로 변동이 크다는 것을 알 수 있다. 이는 특정 용어들이 비슷한 가중점수를 주어 개체를 구별함에 있어서 어려움이 작용했기 때문이다. 따라서 특정 개체에 대하여 낮은 정확률과 재현율을 보이고 있기 때문에 분류하는 데 있어서 좋은 가중치 부여 방법이 아니라 할 수 있다.

하지만 M6 방법에서 다른 방법들에 비교하여 개별 범주에 대해 대부분의 지표들의 분류 결과가 높게 측정되었다. 이는 2.2.3절의 예시와 같이 특정용어들에 대해 IDF, ICSDF 값들이 비슷한 가중점수를 부여받아 개체를 구분하는 능력이 IGM에 비해 떨어지기 때문이다. 따라서 A개체와 C개체에 대하여 완벽한 정확률을 보이고 B개체에 대하여 완벽한 재현율을 보이고 모든 지표에서 가장 작은 표준편차 값을 가지기 때문에 M6 방법이 가장 안정적이라 할 수 있다. 다음으로 전체 범주에 대한 분류 결과를 살펴보기 위해 매크로평균 F_1점수와 마이크로평균 F_1점수를 매 시행마다 계산하여 평균을 구한다. 이 평균을 M1–M6 방법에 따라 계산한 결과를 Figure 3.1에 나타낸다.

Figure 3.1에서 점선으로 연결된 부분은 M1–M6 방법에 대한 매크로평균 F_1점수의 평균값이고 실선으로 연결된 부분은 M1–M6 방법에 대한 마이크로평균 F_1점수의 평균값이다. 마찬가지로 지역적 가중치 함수를 빈도로 이용한 M1–M3 방법보다 지역적 가중치 함수를 빈도에 제곱근을 이용한 M4–M6 방법이 F_1점수가 높음을 알 수 있다. 전역적 가중치 함수 중에는 IGM을 이용한 M3, M6 방법이 가장 F_1점수 값들이 높음을 알 수 있다. IDF와 IDF-ICSDF는 지역적 가중치 함수를 빈도로 이용했을 때는 IDF-ICSDF가 높게 나오지만 빈도에 제곱근을 이용했을 때는 IDF가 높게 나타나므로 두 전역적 가중치 함수는 개체 분류를 하는 데 있어서 IGM에 비해 상대적으로 좋은 방법이 아니라 판단된다.

전체적으로 매크로평균 F_1점수와 마이크로평균 F_1점수 값은 M6 > M4 > M3 > M5 > M2 > M1 순으로 나타났다. 두 F_1점수 값들은 모든 가중치 부여 방법에 대해 비슷한 경향을 보임을 알 수 있고 마이크로평균 F_1점수가 매크로평균 F_1점수가 크게 나타나므로 빈도의 수가 많은 범주에 영향을 받았음을 알 수 있다. 그러나 M6 방법에 대해서 매크로평균 F_1점수는 0.9871, 마이크로평균 F_1점수는 0.9886로

두 점수의 차이가 거의 없기 때문에 용어 가중치 함수로 RTF-IGM 방법을 이용했을 때 개체 내에 빈도의 차이에 영향을 받지 않고 개체 분류를 하는 데 있어서 가장 좋은 가중치 부여 방법이라 판단된다.

4. 결론

문서-용어 빈도행렬에서 단순히 빈도가 높은 용어를 핵심어로 인식하여 개체를 분류하고자 할 경우, 문서와 개체들의 특수성과 대표성이 있는 용어들을 반영하지 못하기 때문에 텍스트를 분류하는 데 있어서 좋은 결과를 얻을 수 없다. 본 연구에서는 이러한 문제를 극복하기 위해 6가지 용어 가중치 함수를 이용하여 문서-용어 가중행렬을 생성하고 핵심어를 추출하는 방법을 고찰하였다. 그리고 2016년 한 해 동안 정부출연연구기관에서 발간한 정기간행물을 수집하고 이를 자료로 활용하여 6개의 문서-핵심어 가중행렬을 생성한 후 문서 분류에서 가장 많이 활용된 SVM을 적용하였다.

SVM 모형의 분류 분석 결과로 모든 가중치 부여 방법 중에서 M6 방법이 매크로평균 F_1점수와 마이크로평균 F_1점수가 다른 방법에 비해 가장 높은 값을 보였다. 따라서 개체 정보가 존재하는 텍스트 자료를 분류하는 데 있어서 지역적 가중치 함수를 빈도에 제곱근을 적용하고 전역적 가중치 함수를 IGM으로 적용하는 문서-용어 가중행렬 생성 방법을 제안한다. 본 연구에서 제안한 방법을 개체 정보가 존재하는 문서 분류를 하는 목적에 활용을 한다면 개체를 대표하는 용어에 더욱 효과적인 가중점수를 부여한 데이터를 생성함을 기대할 수 있다.

References

Chen, K. and Zong, C. (2003). A new weighting algorithm for linear classifier. In *Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering*, 650–655.

Chen, K., Zhang, Z., Long, J., and Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification, *Expert System with Applications*, **66**, 245–260.

Cho, S. G., Cho, J. H., and Kim, S. B. (2015). Discovering meaningful trends in the inaugural addresses of United States Presidents Via text mining, *Journal of Korean Institute of Industrial Engineers*, **41**, 453–460.

Dumais, S. (1991). Improving the retrieval of information from external sources, *Behavior Research Methods, Instruments & Computers*, **23**, 229-236.

Hornik, K., Meyer, D., and Karatzoglou, A. (2006). Support vector machines in R, *Journal of Statisticcal Software*, **15**, 1–28.

Jung, M.J. (2017). *A study on clustering methods for proximity data in text mining* (Master thesis), Pusan National University.

Lee, M. R. and Bae, H. K. (2002). Design of keyword extraction system using TFIDF, *The Korean Society for Cognitive Science*, **13**, 1–11.

Miner, G., Elder, J., and Hill, T. (2012). *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Academic Press, Seoul.

Nakov, P., Popova, A., and Mateev, P. (2001), Weight functions impact on LSA performance. In *Proceeding of the Recent Advances in Natural language processing*, Bulgaria, 187–193.

Ren, F. and Sohrab, M. G. (2013). Class-indexing-based term weighting for automatic text classification, *Information Sciences*, **236**, 109–125.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a "kneedle" in a Haystack: Detecting Knee Points in System Behavior, *Distributed Computing Systems Workshops (ICDCSW) 2011 31st International Conference on, IEEE*, 166–171.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the ACM SIGIR Conference on Research and Development in International Retrieval*, 42–49.

문서 분류를 위한 용어 가중치 기법 비교

정호영[a] · 신상민[b] · 최용석[a,1]

[a]부산대학교 통계학과, [b]동아대학교

(2019년 1월 3일 접수, 2019년 2월 8일 수정, 2019년 2월 12일 채택)

요 약

문서-용어 빈도행렬은 텍스트 마이닝에서 분석하고자 하는 개체 정보를 가지고 있는 일반적인 자료 형태이다. 본 연구에서 문서 분류를 위해 문서-용어 빈도행렬에 적용되는 기존의 용어 가중치인 TF-IDF를 소개한다. 추가하여 최근에 알려진 용어 가중치인 TF-IDF-ICSDF와 TF-IGM의 정의와 장단점을 소개하고 비교한다. 또한 문서 분류 분석의 질을 높이기 위해 핵심어를 추출하는 방법을 제시하고자 한다. 추출된 핵심어를 바탕으로 문서 분류에 있어서 가장 많이 활용된 기계학습 알고리즘 중에서 서포트 벡터 머신을 이용하였다. 본 연구에서 소개한 용어 가중치들의 성능을 비교하기 위하여 정확률, 재현율, F1-점수와 같은 성능 지표들을 이용하였다. 그 결과 TF-IGM 방법이 모두 높은 성능 지표를 보였고, 텍스트를 분류하는데 있어 최적화 된 방법으로 나타났다.

주요용어: 용어 가중치, 문서 분류, 텍스트 마이닝, TF-IDF, 핵심어 추출

[1]교신저자: (46241) 부산시 금정구 부산대학로 63번길 2, 부산대학교 통계학과. E-mail: yschoi@pusan.ac.kr