「정보시스템연구」 제23권 제3호
한국정보시스템학회
2014년 9월, pp. 127~142

http://dx.doi.org/10.5859/KAIS.2014.23.3.127

# 연관분석을 이용한 효과적인 표절검사 및 문서분류에 관한 연구

황 인 수*

<목 차>



## Ⅰ. 서 론

인터넷이 기업의 경영활동뿐만 아니라 개인의 일상생활에 기반이 되고 웹 콘텐츠가 폭발적으로 증가됨에 따라, 인터넷은 신문, 방송, 잡지 등 대중매체로부터 개인의 일상을 기술한 일기에 이르기까지 모든 정보의 원천이 되고 있으며, 정보량의 증가는 정보의 공유를 통해 새로운 가치를 창조하고 있다. 그러나 정보의 공유가 긍정적인 효과만을 가져온 것은 아니며, 표절(plagiarism)과 저작권 침해라는 새로운 사회적 문제를 발생시키고 있다. 표절은 다른 사람의 글을 취하여 마치 자기가 쓴 것처럼 행세하는 행위 등을 일컫는 말로서, 「저작권표준용어집」에서는 "표절이란 다른 사람의 저작물의 전부나 일부를 그대로 또는 그 형태나 내용에 다소 변경을 가하여 자신의 것으로 제공 또는 제시하는 행위"로 정의하고 있다.

표절이 사회적 문제로 대두됨에 따라, 구글(Google)은 상습적으로 저작권을 위반하는 사이트의 검색결과 순위를 낮추도록 검색알고리즘을 개선하였으며, 미국 하원은 인터넷에서 저작권이 침해된 경우 법적으로 웹사이트의 접속을 차단하는 등의 조치를 취할 수 있도록 하는 온라인 저작권침해 금지법안(Stop Online Piracy Act, SOPA)을 제출하였다(Perez, 2012).

* 황인수, 전주대학교 스마트미디어학부 교수, insoo@jj.ac.kr

표절은 저작권 침해라는 법적인 문제를 발생시킬 뿐만 아니라 정보화시대에서 정보의 바람직한 유통 및 활용에 악영향을 미치기 때문에, 표절을 효과적으로 검출하기 위한 연구가 활발히 진행되고 있다.

연구의 방향은 크게 두 가지로 구분되는데, 첫째는 연구논문이나 신문기사 등의 저작권 침해여부를 검증하기 위한 중앙관리시스템을 구축하는 것이며, 둘째는 표절로 의심되는 문서 간의 유사도를 효과적으로 분석하는 것이다(Liu *et al*., 2007). 문서 간의 유사도를 계산하는 방법에는 문장에서 인접한 N개의 음절을 추출하여 상호 비교하는 N-gram 방식, 문장의 변형을 고려하여 문장 내 일부 문자열을 비교하는 문자열 비교 방식, 색인어를 추출하여 가중치($tf$-$idf$)를 부여하는 벡터공간모델, 그리고 단어와 문서 간의 행렬로부터 의미적 관계정보를 추출하는 LSA(Latent Semantic Analysis 방식 등이 있다(지혜성.조준희.임희석, 2010). 그러나 N-gram, 문자열 비교, 그리고 벡터공간모델은 단어나 어절에 의미를 부여하지 못하여 우연히 동일한 단어가 사용된 경우에도 표절로 인식하는 단점이 있으며, LSA 방식은 문장에 사용된 단어에 의존하기 때문에 의미적 관계를 파악하기에 한계가 있으며 표절하지 않은 문서를 표절로 인식할 가능성이 있다.

이에 따라 본 연구는 문장에 사용된 단어 간의 관계를 통해 단어에 의미를 부여하고, 우연히 동일한 단어가 사용된 경우 표절검사에서 제외하기 위해 데이터마이닝의 연관분석기법을 도입하는 방안을 제안한다. 즉, 문장으로부터 추출한 형태소(말뭉치, corpus)를 상품으로 간주하고 문장을 이루는 형태소들로 장바구니(basket)를 구성한 후, 문서 간에 장바구니를 비교 분석하는 것이다. 본 논문의 구성은 다음과 같다. 먼저 제 2장에서는 유사도 결정 방법 및 연관분석에 대해 기술하며, 3장에서는 본 연구에서 제안하는 표절검사 및 문서분류를 위한 연관분석 알고리즘에 대해 기술한다. 4장에서는 알고리즘의 적용사례를 기술하며, 5장에서는 본 연구의 성과를 정리한 후 향후 연구방향을 제시한다.

# Ⅱ. 관련 연구

## 2.1 유사도 계산 방식

문서의 유사도를 계산하는 방식에는 N-gram 방식, 문자열 비교방식, 벡터공간모델 방식, LSA 방식 등이 있다(지혜성.조준희.임희석, 2010).

### 2.1.1 N-gram

N-gram은 문장에서 인접한 N개의 음절을 말하는 것으로서, N-gram 방식은 문장으로부터 N-gram을 추출한 후 문장 간 비교를 통해 유사성을 판단하는 표절검사 방식이다(Paul, 2002). N-gram 방식에서는 문서 내 각 문장에서 빈칸, 마침표, 쉼표 등에 따라 단어 또는 어절을 구분한 후, 각 어절들로부터 N-gram을 추출한다. 예를 들어, “표절검사”라는 어절에서 bi-gram(2문자)은 “표절”, “절검”, “검사”가 되며, tri-gram(3문자)은 “표절검”, “절검사”가 N-gram으로 추출된다. 이 때, 어절의 글자가 N보다 작은 경우에는 전체 어절을 하나의 N-gram으로 간주한다.

어절에 오탈자가 있는 경우에도 유사한 문장으로 판정하는 장점이 있으나, N-gram 방식은 문장의 길이가 길어질 경우 생성되는 N-gram의 개수가 기하급수적으로 증가하기 때문에 많은 저장공간과 처리시간을 필요로 하며, 관련 없는 어절이 N-gram에 의해 일치되는 등의 단점이 있다.

#### 2.1.2 문자열 비교

문자열 비교방식은 전체 문장이 아닌 문장 내 일부 문자열이 비교 대상 문장에 포함되어 있는지를 검사하는 것으로, 문장의 일부 단어를 삭제, 추가, 혹은 변형시킨 경우에도 표절로 검출할 수 있다는 장점이 있다. 그러나 몇 개 단어가 일치되어야 표절로 인정할 것인가에 대한 논란이 있을 수 있는데, 우리나라 교육과학기술부에서는 2008년도에 제정한 표절 가이드라인에서 “6단어 이상 연쇄표현이 일치하는 경우”를 표절 판정의 기준으로 제시하였다. 문자열 비교방식이 문장의 일부분을 변형한 경우에는 유연하게 대응하여 표절을 검출하지만, 어순을 변경하거나 새로운 단어를 삽입 혹은 삭제한 경우에는 표절을 검출하지 못하는 단점이 있다.

#### 2.1.3 벡터공간 모델

벡터공간 모델은 정보검색에서 사용하는 벡터공간 모델을 표절검사에 응용한 것으로, 표절검사 대상 문장으로부터 색인어를 추출하여 벡터로 표현한다(Narayanan, 1995). 색인어에 가중치를 부여하는 방법으로는 $tf\text{-}idf$ 방식을 주로 사용하며, 유사도를 계산하는 방법에는 다이스 유사계수, 자카드 유사계수, 내적 계수, 코사인 유사계수 등이 있는데 일반적으로 코사인 유사계수를 가장 많이 사용한다. 벡터공간 모델은 계산과정이 단순하고, 단어의 배열 순서에 무관하며, 정규화된 유사도 값을 얻을 수 있다는 장점이 있다. 그러나 색인어에 기반하여 표절을 검사하기 때문에 우연히 일치하는 색인어를 표절로 인식할 뿐만 아니라, 비슷한 의미를 갖고 있는 다른 색인어로 변형한 경우에는 표절을 검출하지 못하는 단점이 있다.

#### 2.1.4 LSA(Latent Semantic Analysis)

LSA 방식은 문장에 사용된 단어의 의미관계를 분석하여 표절을 검출하는 방식으로 문서의 의미관계를 찾기 위해 통계 및 선형대수를 이용한다(Georgina, 2012). LSA는 각 문서의 문장에서 단어를 추출하여 단어벡터와 문서벡터로 이루어진 행렬(matrix)로 구성한 후, 특이 값 분해(Singular Vector Decomposition, SVD)를 통해 단어-단어, 단어-문서, 문서-문서로 이루어진 3가지의 행렬을 형성한다. 단어-단어 행렬은 두 개의 단어가 동시에 출현하는 것을 의미하는 공기(co-occurrence)패턴 정보를 포함하고 있다. 이것은 LSA에서 언어의 개념적인 유사정보로써, 각각의 점들로 이루어진 단어벡터나 문서벡터들의 코사인 유사도를 계산하여 구할 수 있다(이동욱 등, 2012). LSA는 단어의 의미적 관계를 이용한다는 장점은 있지만, 문장에서 단어를 추출하여 의미적 유사도를 측정하기 때문에 단어의 개수가 한정적이며, 유사성이 낮은 문장이 의미적 관계에 따라 표절로 판정되는 등 정확도가 낮아지는 단점이 있다(지혜성.조준희.임희석, 2010).

## 2.2 연관분석

### 2.2.1 연관분석의 개념

연관분석은 (Agrawal *et al*., 1993)이 제안한 데이터마이닝 기법의 하나로서, 장바구니 분석을 통한 상품추천 등에 광범위하게 사용되고 있다. 연관분석은 거래내역을 분석하여 각 거래에 동시에 포함되는 상품들을 연관규칙으로 표현하는 데, 연관규칙에 관한 기존의 연구를 간략히 정리하면 다음과 같다.

정의1) $K$개의 상품(product)으로 구성된 집합을 $P = \{p_1, p_2, ...., p_K\}$라고 하고, $P$로부터 임의로 $N$개의 상품을 선택하여 구성한 집합을 $B = \{b_1, b_2, ...., b_N\}$이라고 하면, $B \subseteq P$ 가 되는데, 여기서 $B$을 장바구니(basket)라고 한다.

정의2) $p_1 \rightarrow p_2$의 연관규칙이 존재하기 위해서는 다음의 조건을 만족해야 한다.

① $N$개의 장바구니에서 $p_1$과 $p_2$를 모두 포함하는 장바구니가 $minsupp(\%)$ 이상 존재해야 한다.

② $p_1$을 포함하는 장바구니의 $minconf(\%)$ 이상이 $p_2$를 포함하고 있어야 한다.

위에서 ①은 하나의 상품 혹은 일련의 상품이 전체거래에서 차지하는 비율로 지지도(support)라고 하며, $supp(p_1)$ 혹은 $supp(p_1 \rightarrow p_2)$로 표현한다. 또한 ②는 $p_1$을 포함하는 거래에 $p_2$가 동시에 포함되는 조건부 확률로 신뢰도(confidence)라고 하며, $supp(p_1 \rightarrow p_2)/supp(p_1)$을 의미하는 $conf(p_1 \rightarrow p_2)$로 표현한다.

정의3) 상품 $p_1$과 $p_2$간의 향상도(lift) 또는 관심도(interest)는 상품간의 영향력을 나타나며 다음과 같이 계산한다(Silverstein *et al.* 1998).

$$I(p_1, p_2) = \frac{supp(p_1 \rightarrow p_2)}{supp(p_1)supp(p_2)} = \frac{p(p_1 p_2)}{p(p_1)p(p_2)}$$

향상도는 상품 $p_2$가 $p_1$과 함께 판매된 거래와 상품 $p_2$가 $p_1$에 관계없이 단독으로 판매된 거래의 비율이다. 따라서 향상도가 1이라는 것은 $p_2$가 $p_1$에 영향을 받아서 판매되는 비율과 $p_2$가 $p_1$과 관계없이 단독으로 판매되는 비율이 같음을 의미하기 때문에 이들 두 상품은 서로 독립이다. 또한, 향상도가 1보다 크면 교차판매의 효과를 있음을 의미하기 때문에 보완재의 성격을 가지며, 향상도가 1보다 작으면 $p_1$의 판매가 오히려 $p_2$의 판매를 감소시키기 때문에 대체재의 성격을 갖는다(Brijs *et al*. 1999).

### 2.2.2 단어의 연관성

단어의 연관성은 각 단어의 개별적인 출현빈도와 동시 출현빈도를 이용하여 계산하는 연관계수로 측정한다. 단어의 동시 출현빈도만을 연관계수로 사용하면, 많이 사용되는 단어일수록 동시에 나타나는 빈도가 높아져서 높은 연관계수를 갖게 되므로 연관계수의 값을 상대적으로 비교하기가 어렵다(정영미.이재윤, 1998). 따라서 단어의 개별적인 출현빈도와 전체 단어의 빈도를 함께 이용하여 단어의 통계적인 연관성을 객관적으로 평가하는 상대 동시출현빈도 방식의 연관계수를 주로 사용한다. 연관계수를 측정하는 방법에는 여러 가지가 있으나, 주로 정보이론에 근거하는 상호정보량과 상대 엔트로피를 사용한다. 먼저 상호정보량은 두 독립사건

의 확률변수 $x$와 $y$ 사이의 의존관계를 정량적으로 나타낸 것으로서 다음 계산식에 따라 계산한다.

$$M(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

상호정보량은 두 확률이 완전히 독립일 경우 0이 되며, 의존관계가 깊을수록 큰 값은 갖는다. 또한 대칭성을 만족하므로 $M(x,y) = M(y,x)$가 성립한다. 연관성 분석에서 상호정보량을 이용할 때의 문제점으로는 빈도가 낮은 단어의 상호정보량이 빈도가 높은 단어의 상호정보량보다 상대적으로 과대평가되는 경향이 있다. 상대 엔트로피는 두 확률분포 $p(x)$와 $q(x)$간의 평균적인 차이를 측정하는 것으로서, $KL$거리(Kullback-Leiber Distance: $D_{KL}$) 혹은 교차 엔트로피(cross entropy)라고도 한다(정석경, 1997). 상대 엔트로피는 다음 계산식에서 보는 바와 같이, 항상 0보다 크거나 같으며 두 확률이 일치할 경우에만 0이 되고 대칭성을 만족하지 않는다.

$$D_{KL}(p(x) \parallel q(x)) = \sum_x p(x) \left[ \log \frac{1}{q(x)} - \log \frac{1}{p(x)} \right]$$

# Ⅲ. 연관분석을 이용한 문서유사도 계산

## 3.1 형태소 분석

앞에서 설명한 N-gram 방식은 문장으로부터 N개의 어절을 추출하여 비교하는 것으로서, "문서유사도"에 2-gram을 적용할 경우 차례대로 2글자씩을 묶어서 "문서, 서유, 유사, 사도"를 어절로 추출하여 비교한다. 이 방식은 많은 수의 N-gram이 생성될 뿐만 아니라, 위의 예에서 제자를 의미하는 "사도(disciple)"와 같이 의도하지 않은 어절을 비교하게 되는 문제가 있다. 따라서 검사의 정확도를 높이기 위해서 형태소분석을 통해 추출된 문자열을 단어 혹은 어절로 사용하는 것이 바람직하다. 본 연구에서는 검색엔진 오픈소스 프로젝트 루씬(lucene)에 결합하여 사용할 수 있도록 JAVA 기반으로 개발된 "루씬 한글 분석기"를 이용하여 형태소를 분석했다. 이것은 MorphAnalyzer의 analyze 메소드를 통해 분석을 진행하며, 분석결과를 명사(N), 동사(V) 등으로 구분하여 List 구조로 반환하고, 복합명사를 단위명사로 구분하는 기능도 갖고 있다. 본 연구에서는 형태소분석을 통해 문서 내의 각 문장에서 추출한 단위명사(N)를 이용하여 연관분석을 실시하였다.

## 3.2 문서의 유사도

<그림 1>은 문서 간 유사도를 계산한 후 유사도에 따라 문서를 분류하는 과정을 설명하고 있다. 문서의 유사도를 계산하기 위해서는 먼저 형태소 분석을 통해 추출한 형태소를 이용하여 형태소 쌍을 구성하는데, 하나의 문장에서 추출한 형태소들이 연관분석의 장바구니를 구성하기 때문에 형태소 쌍은 문장단위로 구성한다. 이 때 하나의 쌍에 포함되는 형태소의 개수가 많아질수록 분석에 필요한 데이터양이 기하급수적으로 증가하기 때문에 본 연구에서는 2개 혹은 3개의 형태소가 쌍을 이루도록 하였다. 참고로, 총 10개의 형태소를 갖는 문장에서 형태소를 2개씩 쌍으로 묶을 경우 ${}_{10}C_2 = 45$개의 쌍이 생성되며, 3개씩 쌍으로 묶을 경우

${}_{10}C_3 = 120$개의 쌍이 생성된다.

다음으로 문서에서 각 형태소 쌍이 출현하는 빈도를 계산한 후, 계산의 복잡도를 줄이기 위해 출현빈도가 F보다 작은 형태소 쌍은 분석에

표절검사를 실시할 파일을 읽어서 문서집합을 생성한다.
각 문서를 문장단위로 세분화한 후,
  형태소분석기를 이용하여 각 문장으로부터 형태소를 추출하고,
  형태소 쌍(연관관계)을 생성한다.
형태소 쌍들이 각 문서에 출현하는 빈도로부터 지지도를 계산한 후,
  출현빈도가 F개 미만인 형태소 쌍을 제거하고,
  문서 내 형태소 쌍의 지지도 합이 1이 되도록 정규화 한다.
문서 간에 형태소 쌍을 비교하여
  일치하는 형태소 쌍의 지지도를 합산하여 문서 간 유사도를 계산한다.

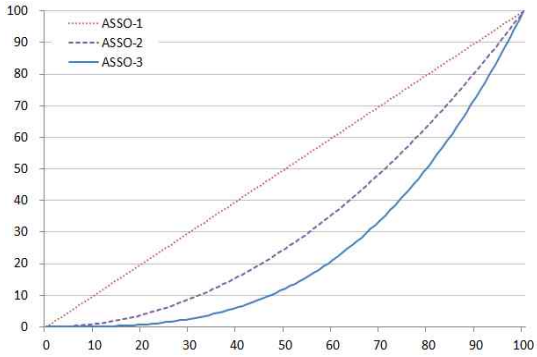
자기 자신을 제외한 다른 모든 문서들과의 평균유사도($A_T$)를 계산한다.
모든 문서가 그룹에 할당될 때까지 다음을 반복한다.
  새로운 그룹($G_i$)을 생성한다.
  남아 있는 문서가 한 개이면,
    이 문서를 앞에서 생성한 새로운 그룹($G_i$)에 추가한다.
  그렇지 않으면,
    상호간 유사도($A_{jk}$)가 가장 큰 두 개의 문서 $j, k$를 선택한 후,
      상호간 유사도($A_{jk}$)가 전체 평균유사도($A_T$)보다 크면,
        선택된 두 개의 문서 $j, k$를 모두 그룹($G_i$)에 추가한다.
      그렇지 않으면,
        그룹($G_i$)이 첫 번째 그룹이면,
          다른 모든 문서들과의 평균유사도($A_{dT}$)가 가장 큰 문서 $d$를 선택하여 그룹($G_i$)에 추가한다.
        그렇지 않으면
          직전 그룹($G_{i-1}$)내의 문서들과의 유사도평균($A_{di-1}$)이 가장 큰 문서 $d$를 선택하여 직전 그룹($G_{i-1}$)에 추가한다.
  다음을 반복한다.
    그룹($G_i$)내 문서들과의 평균유사도($A_{dG_i}$)가 가장 큰 문서 $d$를 선택한 후,
      그룹($G_i$)내 문서들과의 평균유사도($A_{dG_i}$)가 전체 평균유사도($A_T$)보다 크거나 같으면,
        문서 $d$를 그룹($G_i$)에 추가한다.
      그렇지 않으면,
        그룹($G_i$)에 대한 문서할당을 종료한다.
각 그룹($G_i$)에 대해
  각 문서 $d$를 그룹내 문서들과의 평균유사도($A_{dG_i}$)에 따라 내림차순으로 정렬한다.

<그림 1> 문서의 유사도계산 및 문서분류 과정

서 제외한다. 본 연구에서는 시뮬레이션을 통해 N을 2로 설정하였으나, 문서의 길이가 짧은 경우에는 대부분의 형태소 쌍이 제외될 수 있으므로 모든 형태소 쌍을 포함하도록 F를 1로 설정할 수도 있다. 문서의 길이에 따른 영향력을 최소화하기 위해 각 형태소 쌍의 지지도(support)의 합이 1이 되도록 지지도를 정규화한다.

각 문서에 대해 형태소 쌍을 구성하고 지지도 계산이 완료되면, 문서 간에 일치하는 형태소 쌍에 따라 문서 간 유사도를 계산한다. 문서 A의 문서 B에 대한 유사도는 문서 B에 존재하는 문서 A의 형태소 쌍들의 지지도 합으로 계산되며, 문서 B의 문서 A에 대한 유사도는 문서 A에 존재하는 문서 B의 형태소 쌍들의 지지도 합으로 계산된다. 따라서 문서 간 유사도가 비대칭이 되기 때문에, 문서 간 표절의 방향을 예측할 수도 있다.

<그림 2>는 본 연구에서 사용하는 유사도 함수 곡선을 보여주고 있다. 그림에서 *ASSO-1*은 문서간의 유사도를 계산할 때 $tf\text{-}idf$ 방식에 따라 개별단어를 사용하기 때문에 각 단어의 가중치가 동일할 경우 문서 간에 일치하는 단어의 비율에 비례하여 유사도가 결정된다. 그러나 *ASSO-2*와 *3*은 문서간의 유사도를 계산하기 위해 본 연구에서 제안하는 2개 또는 3개의 형태소 쌍을 이용하여 유사도를 계산하기 때문에 지수함수를 나타낸다. 예를 들어, 형태소별로 가중치가 동일하다는 가정하에, 10개의 형태소를 갖는 문장에서 1개의 형태소가 일치할 경우 $tf\text{-}idf$ 방식에서의 유사도는 10%가 되지만, *ASSO-2,3*에서는 형태소쌍이 생성되지 않으므로 유사도는 0%가 된다. 또한, 10개의 형태소




<그림 2> 유사도 함수 곡선

중에서 9개의 형태소가 일치할 경우 $tf\text{-}idf$ 방식에서의 유사도는 90%가 되지만, *ASSO-2*에서는 총 45개의 형태소 쌍 중에서 9개 쌍이 불일치하므로 유사도는 80%에 불과하게 된다.

이러한 결과는 일반적인 판단기준에 의한 문서 간의 유사도와 상이할 결과를 도출할 수도 있으나, 일치하는 형태소의 개수가 증가할수록 유사도를 높게 계산한다는 점에서 유의한 함수가 될 수 있을 것으로 판단된다. 실제 시뮬레이션 결과에서도 유사도 계산 및 문서분류에서 좋은 성과를 나타냈기 때문에, 본 연구에서는 유사도 함수를 보정하지 않고 그대로 적용하였다.

### 3.3 문서 분류

문서 간의 유사도는 표절검사 뿐만 아니라, 문서의 내용에 따른 문서분류에도 사용될 수 있다. 특히 본 연구에서는 문서 간 유사도가 비대칭으로 나타나기 때문에, 그룹 내에서 주요문서를 추천하는 것도 가능하게 된다. 본 연구에서 그룹을 형성하는 과정은 앞의 <그림 1>과 같으며, 이를 간략히 설명하면 다음과 같다.

문서 $d$를 그룹($G_i$)에 포함시킬 것인지의 여부를 판단하는 기준으로 각 문서간의 전체 평균유사도($A_T$)를 사용한다. 즉, 특정 문서 $d$를 그룹($G_i$)에 포함시키고자 할 때, 문서 $d$와 그룹 내의 다른 문서들과의 유사도($A_{dG_i}$)가 전체 평균유사도($A_T$)보다 크거나 같으면 그룹($G_i$)에 포함시키고 그렇지 않으면 포함시키지 않는 것이다. 결과적으로 유사도가 큰 문서들은 그룹에 포함되고 그렇지 않은 문서들을 그룹에서 제외됨으로써, 그룹 내 문서 간의 유사도는 그룹 외 문서 간의 유사도보다 커지게 된다. 새로운 문서그룹을 형성할 때는 아직까지 그룹에 포함되지 않고 남아있는 문서 중에서 문서 간의 유사도가 가장 큰 2개의 문서 $j, k$를 선택하여 그룹을 형성한다. 그러나 문서 간의 유사도($A_{jk}$)가 전체 평균유사도($A_T$)보다 작은 경우에는, 그룹($G_i$)가 첫 번째 그룹이면 다른 모든 문서들과의 평균유사도($A_{dT}$)가 가장 큰 문서 $d$를 선택하여 그룹($G_i$)에 추가하고, 그렇지 않으면 직전 그룹($G_{i-1}$)내의 문서들과의 유사도평균($A_{di-1}$)이 가장 큰 문서 $d$를 선택하여 직전 그룹($G_{i-1}$)에 추가하여 그룹을 형성한다.

새로운 그룹$G_i$)이 형성되면, 그룹 내의 문서들과의 유사도가 평균유사도($A_{dG_i}$)보다 큰 문서들을 유사도 크기에 따라서 차례대로 그룹에 추가한다. 그룹에 포함되지 못하고 남아있는 문서들은 위 그룹과 이질적인 문서로서, 새로운 그룹을 형성하도록 위의 과정을 반복한다. 모든 문서에 대한 그룹핑 과정이 완료되면, 각 그룹별로 그룹 내 문서 간 유사도($A_{dG_i}$)에 따라 정렬한다. 여기서, 그룹 내 문서 간의 유사도가 가장 큰 문서는 다른 문서들과의 공통점을 가장 많이 갖고 있는 문서로서 해당 그룹의 핵심문서(Key Document)가 된다.

# Ⅳ. 적용사례 분석

## 4.1 자기소개서

첫 번째 사례는 입학사정관제 전형에 지원한 학생들이 작성한 자기소개서의 표절여부를 검사하는 것이다. 자기소개서는 주어진 질문에 따라 작성하기 때문에 서로 다른 자기소개서에 동일한 단어들이 많이 사용될 수 있다. 또한, 500 글자 이내로 작성하기 때문에 형태소 쌍의 출현빈도가 높지 않으므로 출현빈도가 1인 형태소 쌍을 포함하여 모든 형태소 쌍을 분석 대상으로 하였다.

<표 1> tf-idf방식의 문서 간 유사도

| To / From | A | B | C | D | E |
|---|---|---|---|---|---|
| **A** | **100.0** | **78.5** | 7.1 | 11.0 | 13.7 |
| **B** | **75.8** | **100.0** | 6.9 | 9.2 | 14.9 |
| **C** | 7.2 | 6.6 | **100.0** | **21.3** | 7.1 |
| **D** | 11.9 | 10.3 | **22.5** | **100.0** | 7.4 |
| **E** | 14.7 | 16.1 | 6.6 | 7.3 | **100.0** |

<표 2> 연관분석기준 문서 간 유사도

| To / From | A | B | C | D | E |
|---|---|---|---|---|---|
| **A** | **100.0** | **59.1** | 0.2 | 0.2 | 0.2 |
| **B** | **51.0** | **100.0** | 0.0 | 0.0 | 0.2 |
| **C** | 0.2 | 0.0 | **100.0** | 1.4 | 0.0 |
| **D** | 0.1 | 0.0 | 0.9 | **100.0** | 0.0 |
| **E** | 0.2 | 0.2 | 0.0 | 0.0 | **100.0** |

<표 1>은 본 연구에서 제안하는 형태소의 쌍을 이용하는 대신 tf-idf 방식과 같이 개별 단어

를 이용하여 유사도를 계산한 결과로서, 자기소개서 A와 B, 그리고 C와 D가 각각의 그룹을 형성하였다. 그러나 수작업으로 확인한 결과 C와 D는 표절을 하지 않은 것으로 나타났는데, 이것은 자기소개서가 주어진 질문에 따라 작성되기 때문에 자기소개서 간에 동일한 단어를 많이 포함하고 있어서 20% 이상의 유사도를 나타냈던 것이다. 다음으로, <표 2>는 연관분석을 적용한 결과를 보여주고 있다. 즉, 자기소개서의 각 문장으로부터 형태소 쌍을 추출한 후, 자기소개서 간에 형태소 쌍을 비교하여 유사도를 측정한 것으로서, 자기소개서 A와 B만 그룹을 형성하였다. C와 D간의 유사도는 약 1% 수준으로 현저히 줄어들었음을 볼 수 있는데, 이것은 전체적으로는 동일한 단어들이 많이 사용되었지만 각 문장은 표절되지 않았음을 보여주는 결과이다.

## 4.2 신문기사

기사는 사실을 전하기 위해 작성하는 글로서, 주어진 사실을 그대로 기술하거나 혹은 약간의 의견을 부가하여 작성한다. 따라서 기사 간에는 일반적으로 높은 유사도를 보이며 때로는 표절로 의심되는 경우도 많다. 뿐만 아니라, 서로 다른 주제 및 관점을 갖고 쓴 기사인 경우에도, 관련분야의 기사는 유사한 단어를 많이 사용할 수밖에 없어서 기사 간 유사도가 높게 나타난다.

기사를 분석하기 위해 언론의 많은 관심을 받았던 “나로호”를 검색어로 다음(Daum)과 구글(Google)을 검색하였다. 검색결과, 나로호의 미래에 대한 심층분석기사, 청와대에서 나로호 발사를 만류했다는 기사, 나로호 부품에 사용된 슈퍼섬유에 관한 기사, 그리고 나로호 관광에 관한 기사 등으로 구분될 수 있었다. 다음과 구글은 <그림 3>과 <그림 4>에서 보는 바와 같이, 검색결과를 주제에 따라 그룹핑하여 보여준다. 기사의 내용을 검토한 결과, 다음의 광주일보, 경향신문, 시사서울은 별도로 그룹핑 되었으나 상위의 연합뉴스 등에 그룹핑하는 것이 더 바람직하며, 구글의 대덕넷은 서로 관련 없는 기사에 잘 못 그룹핑된 것으로 나타났다.


<그림 3> 다음(Daum)의 〝나로호〞 검색결과

<표 3>은 $tf\text{-}idf$ 방식에 따른 유사도분석결과를 보여주고 있는데, 총 17개의 신문기사가 3개의 그룹과 3개의 독립기사로 분류되었다. 첫 번째 그룹은 중소기업에서 만든 슈퍼섬유에 관한 기사이며, 두 번째 그룹은 청와대에서 나로호 발사를 만류한 기사, 그리고 세 번째 그룹은 나로호 관광을 다루고 있다. 첫 번째 그룹에서 한


〈그림 4〉 구글(Google)의 "나로호" 검색결과

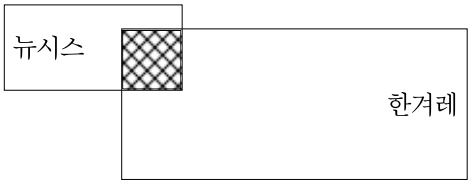
국섬유신문의 기사는 다른 기사와 비교할 때 상대적으로 낮은 유사도를 나타내고 있는데, 이것은 한국섬유신문이 슈퍼섬유에 초점을 맞추어 기사를 작성한 반면에, 다른 신문들은 슈퍼섬유를 중심으로 하되 대구지역의 섬유산업에 초점을 맞추었기 때문으로 기사를 엄격히 분류할 경우에는 서로 다른 그룹에 속할 수도 있다.

모든 신문들이 나로호에 관한 기사를 다루고 있기 때문에 세부그룹에 속하지 않은 기사 간에도 동일한 단어들이 많이 사용되어 약 10~30%의 유사도를 나타내고 있다. 특히 독립적으로 그룹을 형성하고 있는 한겨레와 대덕넷의 기사는 표의 세로 열에서 다른 모든 기사들에 대해 약 10~30%의 비교적 높은 유사도를 나타내고 있는데, 이는 한겨레와 대덕넷이 다른 기사들이 다루고 있는 나로호에 대한 일반적인 현황에 부가하여 추가적인 많은 내용을 담고 있기 때문이다. 반대로, 한겨레와 대덕넷의 가로 행은 15% 미만의 유사도를 나타내고 있는데, 이는 한겨레와 대덕넷에서 다루고 있는 기사들이 다른 기사에서는 많이 다루지 않고 있는 내용임을 알 수 있다. 참고로, 한겨레와 뉴시스의 유사도는 각각 6.9%와 21.1%로 나타나고

<표 3> $tf$-$idf$기준 신문기사간 유사도 분석결과 (그룹지수 59.7%)

| From \ To | 동아① | 사이언스 | 한국섬유 | 연합① | 서울신문 | 동아② | 헤럴드 | 광주일보 | 시사서울 | 경향 | 한강 | 뉴스1 | 연합② | 포커스 | 뉴시스 | 한겨레 | 대덕넷 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 동아① | 100.0 | 100.0 | 35.8 | 12.0 | 12.6 | 11.7 | 9.8 | 12.0 | 12.2 | 9.7 | 9.7 | 7.2 | 6.9 | 5.5 | 6.2 | 18.9 | 19.1 |
| 사이언스 | 100.0 | 100.0 | 35.7 | 12.0 | 12.6 | 11.6 | 9.8 | 12.0 | 12.2 | 9.7 | 9.7 | 7.2 | 6.9 | 5.4 | 6.2 | 18.9 | 19.1 |
| 한국섬유 | 48.9 | 48.9 | 100.0 | 11.7 | 11.7 | 11.7 | 10.1 | 10.6 | 12.0 | 8.3 | 9.5 | 4.6 | 4.6 | 4.3 | 3.2 | 28.2 | 20.6 |
| 연합① | 13.1 | 13.1 | 8.5 | 100.0 | 85.5 | 96.2 | 73.4 | 71.2 | 51.1 | 40.3 | 11.8 | 10.1 | 7.0 | 10.5 | 8.9 | 21.4 | 23.6 |
| 서울신문 | 14.9 | 14.9 | 9.5 | 90.9 | 100.0 | 87.0 | 83.3 | 82.2 | 56.9 | 43.3 | 13.0 | 10.7 | 7.9 | 11.6 | 9.9 | 23.2 | 23.2 |
| 동아② | 12.9 | 12.9 | 8.8 | 97.8 | 83.1 | 100.0 | 70.9 | 70.3 | 50.4 | 39.2 | 11.0 | 10.5 | 7.2 | 9.1 | 8.0 | 21.7 | 24.4 |
| 헤럴드 | 12.6 | 12.6 | 8.1 | 93.4 | 100.0 | 88.3 | 100.0 | 89.5 | 57.3 | 50.2 | 13.4 | 11.0 | 8.1 | 12.4 | 9.0 | 22.1 | 20.7 |
| 광주일보 | 15.1 | 15.1 | 8.7 | 85.1 | 92.4 | 82.1 | 84.5 | 100.0 | 55.1 | 46.3 | 12.9 | 12.2 | 9.3 | 13.9 | 10.3 | 23.6 | 18.8 |
| 시사서울 | 15.4 | 15.4 | 8.5 | 75.3 | 79.2 | 72.4 | 67.5 | 67.5 | 100.0 | 47.8 | 13.9 | 7.5 | 4.4 | 8.8 | 7.7 | 26.1 | 24.6 |
| 경향 | 13.6 | 13.6 | 7.3 | 58.2 | 58.2 | 56.0 | 58.2 | 56.0 | 49.4 | 100.0 | 13.8 | 10.9 | 8.8 | 12.4 | 9.1 | 21.4 | 21.9 |
| 한강 | 11.4 | 11.4 | 5.5 | 11.2 | 11.2 | 9.3 | 10.2 | 9.3 | 11.9 | 8.4 | 100.0 | 39.5 | 35.5 | 41.5 | 10.8 | 11.0 | 13.1 |
| 뉴스1 | 14.7 | 14.7 | 5.1 | 18.4 | 17.0 | 18.4 | 17.0 | 18.4 | 15.6 | 14.3 | 64.0 | 100.0 | 84.9 | 47.3 | 15.2 | 17.3 | 21.1 |
| 연합② | 14.8 | 14.8 | 5.5 | 12.8 | 12.8 | 12.8 | 12.8 | 14.3 | 12.8 | 12.8 | 67.4 | 96.9 | 100.0 | 48.4 | 16.7 | 16.1 | 17.4 |
| 포커스 | 10.9 | 10.9 | 5.6 | 20.3 | 20.3 | 16.7 | 20.3 | 22.0 | 17.4 | 17.9 | 68.4 | 50.2 | 44.4 | 100.0 | 21.0 | 19.4 | 17.0 |
| 뉴시스 | 9.3 | 9.3 | 1.4 | 14.3 | 14.3 | 11.8 | 12.7 | 14.3 | 14.0 | 10.2 | 22.5 | 14.8 | 13.5 | 17.5 | 100.0 | 21.1 | 12.9 |
| 한겨레 | 14.3 | 14.3 | 14.4 | 12.2 | 12.2 | 12.0 | 10.6 | 11.7 | 12.8 | 9.5 | 8.0 | 6.7 | 6.0 | 6.3 | 6.9 | 100.0 | 21.7 |
| 대덕넷 | 13.8 | 13.8 | 12.4 | 14.0 | 13.1 | 14.0 | 10.8 | 11.8 | 11.7 | 9.7 | 8.6 | 7.9 | 6.8 | 6.7 | 6.2 | 21.0 | 100.0 |

있는데, 이는 <그림 5>와 같은 집합구조를 갖고 있기 때문이다.




<그림 5> 신문 기사내용 일치의 예

<표 4>는 연관분석을 이용하여 기사 간의 유사도를 계산한 결과로서, 형태소 2개씩으로 쌍을 구성한 후 계산의 복잡도를 줄이기 위해 형태소 쌍의 출현빈도가 2보다 작은 것은 제외하였다. 시뮬레이션 결과, $tf\text{-}idf$ 방식에서는 독립적인 그룹을 형성하였던 뉴시스가 세 번째 그룹에 포함되는 것으로 나타났는데, 이것은 뉴시스가 자원봉사를 다루면서 세 번째 그룹의 공통주제인 나로호 관광도 일부 기술하고 있기 때문이다.

본 연구에서는 문서분류의 질을 평가하기 위해 문서간 유사도의 총합에서 그룹 내 문서 간 유사도의 합이 차지하는 비율로 계산되는 그룹지수를 도입하였다. 그룹지수를 계산할 때, 표의 대각선에 위치한 자기 자신과의 유사도는 항상 100%이므로 유사도의 총합뿐만 아니라 그룹 내 문서 간의 유사도 계산에서 제외하였다. $tf\text{-}idf$ 방식의 그룹지수는 59.7%에 불과하였으나, 연관분석을 이용한 결과 그룹지수가 83.8%로 현저히 증가하였다. 이것은 $tf\text{-}idf$ 방식은 단어의 존재여부에 따라 유사도를 계산하기 때문에 우연히 사용된 단어 혹은 서로 다른 주제와 관점에서 작성되었으나 해당 분야의 용어가 반복적으로 사용된 경우 문서 간의 유사도가 높아지지만, 연관분석에서는 하나의 문장에서 2개 이상의 단어가 동일하게 사용되어야 유사한 문장으로 인정하기 때문이다.

<표 3>에서는 각 그룹에 속하지 않는 기사, 즉 서로 관련이 적은 기사간의 유사도가 10~

<표 4> 연관분석(ASSO-2)을 이용한 신문기사간 유사도 분석결과 (그룹지수 83.8%)

| From \ To | 동아① | 사이언스 | 한국섬유 | 연합① | 서울신문 | 동아② | 광주일보 | 헤럴드 | 시사서울 | 경향 | 연합② | 한강 | 뉴스1 | 포커스 | 뉴시스 | 한겨레 | 대덕넷 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 동아① | 100.0 | 100.0 | 25.9 | 2.7 | 2.7 | 0.5 | 2.1 | 2.7 | 1.6 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 1.6 | 0.0 | 3.5 |
| 사이언스 | 100.0 | 100.0 | 25.9 | 2.7 | 2.7 | 0.5 | 2.1 | 2.7 | 1.6 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 1.6 | 0.0 | 3.5 |
| 한국섬유 | 23.0 | 23.0 | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 3.8 |
| 연합① | 4.9 | 4.9 | 0.0 | 100.0 | 96.5 | 74.8 | 65.7 | 52.8 | 36.7 | 23.4 | 3.1 | 3.1 | 0.0 | 0.0 | 3.1 | 4.5 | 6.3 |
| 서울신문 | 5.1 | 5.1 | 0.0 | 100.0 | 100.0 | 73.9 | 68.1 | 54.7 | 38.0 | 24.3 | 3.3 | 3.3 | 0.0 | 0.0 | 3.3 | 4.7 | 6.5 |
| 동아② | 1.5 | 1.5 | 0.0 | 78.9 | 75.1 | 100.0 | 48.7 | 42.1 | 27.6 | 23.8 | 2.3 | 2.3 | 1.5 | 0.0 | 3.1 | 4.2 | 3.1 |
| 광주일보 | 5.9 | 5.9 | 0.0 | 100.0 | 100.0 | 72.0 | 100.0 | 66.1 | 38.7 | 34.4 | 3.8 | 3.8 | 0.0 | 0.0 | 4.3 | 2.2 | 6.5 |
| 헤럴드 | 5.1 | 5.1 | 0.0 | 62.1 | 62.1 | 48.2 | 51.8 | 100.0 | 24.5 | 20.9 | 5.5 | 5.1 | 0.0 | 0.0 | 3.6 | 4.7 | 8.3 |
| 시사서울 | 4.3 | 4.3 | 0.0 | 47.8 | 47.8 | 34.8 | 32.9 | 27.1 | 100.0 | 18.4 | 6.3 | 7.2 | 2.4 | 1.0 | 4.8 | 7.2 | 5.8 |
| 경향 | 2.1 | 2.1 | 0.0 | 39.6 | 39.6 | 41.0 | 38.2 | 29.9 | 24.3 | 100.0 | 2.8 | 2.8 | 1.4 | 0.0 | 3.5 | 0.0 | 4.2 |
| 연합② | 0.0 | 0.0 | 0.0 | 3.8 | 3.8 | 3.0 | 2.6 | 4.2 | 7.2 | 1.9 | 100.0 | 40.4 | 22.6 | 15.8 | 14.0 | 12.1 | 6.8 |
| 한강 | 0.0 | 0.0 | 0.0 | 4.3 | 4.3 | 4.3 | 3.4 | 4.3 | 8.2 | 2.9 | 41.5 | 100.0 | 19.8 | 14.0 | 16.4 | 8.7 | 4.8 |
| 뉴스1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.9 | 0.0 | 0.0 | 6.9 | 3.4 | 60.3 | 55.2 | 100.0 | 46.6 | 15.5 | 10.3 | 6.9 |
| 포커스 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 5.7 | 0.0 | 62.9 | 65.7 | 74.3 | 100.0 | 20.0 | 5.7 | 0.0 |
| 뉴시스 | 4.3 | 4.3 | 0.0 | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 | 3.1 | 12.1 | 12.1 | 4.3 | 3.1 | 100.0 | 1.2 | 1.2 |
| 한겨레 | 0.0 | 0.0 | 0.2 | 2.4 | 2.4 | 2.2 | 0.6 | 1.3 | 2.8 | 0.0 | 2.7 | 2.0 | 0.9 | 0.3 | 0.3 | 100.0 | 4.1 |
| 대덕넷 | 0.9 | 0.9 | 0.9 | 1.5 | 1.5 | 0.6 | 0.8 | 2.0 | 1.1 | 0.4 | 1.7 | 0.8 | 0.5 | 0.0 | 0.2 | 4.3 | 100.0 |

<표 5> 연관분석(ASSO-3)을 이용한 신문기사간 유사도 분석결과 (그룹지수 96.7%)

| From \ To | 동아① | 사이언스 | 한국섬유 | 연합① | 서울신문 | 광주일보 | 동아② | 헤럴드 | 시사서울 | 경향 | 연합② | 한강 | 뉴스1 | 포커스 | 뉴시스 | 한겨레 | 대덕넷 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 동아① | 100.0 | 100.0 | 8.1 | 0.5 | 0.5 | 0.5 | 0.0 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 |
| 사이언스 | 100.0 | 100.0 | 8.1 | 0.5 | 0.5 | 0.5 | 0.0 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.0 | 0.0 |
| 한국섬유 | 4.8 | 4.8 | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.5 |
| 연합① | 0.7 | 0.7 | 0.0 | 100.0 | 98.6 | 64.8 | 50.7 | 35.2 | 19.3 | 10.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.7 | 0.0 |
| 서울신문 | 0.7 | 0.7 | 0.0 | 100.0 | 100.0 | 65.7 | 50.0 | 35.7 | 19.6 | 10.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.7 | 0.7 | 0.0 |
| 광주일보 | 1.1 | 1.1 | 0.0 | 100.0 | 100.0 | 100.0 | 44.7 | 47.9 | 21.3 | 15.4 | 0.0 | 0.0 | 0.0 | 0.0 | 1.1 | 0.0 | 0.0 |
| 동아② | 0.0 | 0.0 | 0.0 | 54.8 | 53.3 | 30.7 | 100.0 | 28.0 | 11.5 | 10.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8 | 0.8 | 0.0 |
| 헤럴드 | 0.6 | 0.6 | 0.0 | 32.0 | 32.0 | 28.5 | 23.1 | 100.0 | 7.7 | 5.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.0 | 0.0 |
| 시사서울 | 0.0 | 0.0 | 0.0 | 14.9 | 14.9 | 10.8 | 8.0 | 6.4 | 100.0 | 2.8 | 1.1 | 1.7 | 0.0 | 0.0 | 0.0 | 0.6 | 0.0 |
| 경향 | 0.0 | 0.0 | 0.0 | 17.7 | 17.7 | 17.7 | 19.2 | 11.5 | 7.7 | 100.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 연합② | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.9 | 0.0 | 100.0 | 12.8 | 5.7 | 3.5 | 1.1 | 0.7 | 0.0 |
| 한강 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.7 | 0.0 | 16.8 | 100.0 | 3.7 | 2.8 | 2.3 | 0.0 | 0.0 |
| 뉴스1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 42.9 | 23.8 | 100.0 | 28.6 | 0.0 | 0.0 | 0.0 |
| 포커스 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 55.6 | 44.4 | 66.7 | 100.0 | 0.0 | 0.0 | 0.0 |
| 뉴시스 | 0.6 | 0.6 | 0.0 | 0.6 | 0.6 | 0.6 | 0.6 | 0.6 | 0.0 | 0.0 | 1.2 | 1.2 | 0.0 | 0.0 | 100.0 | 0.0 | 0.0 |
| 한겨레 | 0.0 | 0.0 | 0.0 | 0.5 | 0.5 | 0.0 | 0.5 | 0.0 | 0.5 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 | 0.0 |
| 대덕넷 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 100.0 |

30%로 높게 나타났으나, <표 4>에서는 상당수의 유사도가 0%로 계산되었으며, 나머지도 대부분 10% 미만으로 크게 낮아졌음을 볼 수 있다. 특히 <표 3>에서 21.18%로 나타난 뉴시스의 한겨레에 대한 유사도는 <표 4>에서는 1.2%로 현저하게 감소되었음을 볼 수 있다.

결과적으로, 문서의 표절검사 및 문서분류에 연관분석의 개념을 도입할 경우, 형태소 쌍이 각 형태소에 의미(semantic)을 부가하는 역할을 하기 때문에 표절검사 및 문서분류의 정확도를 크게 향상시킬 수 있다.

2008년도 교육과학기술부의 표절 가이드라인에 따르면 “6단어 이상 연쇄표현이 일치하는 경우”를 표절 판정의 기준으로 삼고 있다. 따라서 본 연구에서는 앞에서 2개의 형태소를 쌍으로 구성하였던 것을 3개로 확대하여 동일한 시뮬레이션을 실시하였다. 시뮬레이션 결과, ASSO-3 방식은 보다 엄격한 기준이 적용됨에 따라 한국섬유신문이 첫 번째 그룹으로부터 분리되고, 뉴시스는 $tf\text{-}idf$ 방식의 결과와 동일하게 세 번째 그룹으로부터 분리되었다. 이것은 앞에서 설명한 바와 같이, 첫 번째 그룹은 대부분이 슈퍼섬유와 섬유산업을 함께 다루고 있는 반면에 한국섬유신문은 슈퍼섬유만을 다루었기 때문이며, 세 번째 그룹은 주로 관광을 다루고 있는 반면에 뉴시스는 자원봉사활동에 초점을 맞추고 있기 때문이다.

앞에서 2개의 형태소들로 쌍을 구성하였을 때에는 그룹지수가 83.8%이었으나, 3개의 형태소들로 쌍을 구성한 결과 그룹지수가 96.7%로 크게 향상되었다. 이것은 그룹에 속하지 않은 신문기사간의 유사도의 대부분은 0%에 가까운 값을 갖기 때문이다. $tf\text{-}idf$ 방식에서 평균 20%의 높은 유사도를 나타냈던 한겨레 및 대덕넷과 다른 문서 간의 유사도도 1% 미만으로 감소하였다. 따라서 본 연구에서 제안하는 연관분석 방법을 이용할 경우, 교육과학기술부의 가이드라인인 6단어 대신 3단어만 사용하더

라도 표절검사 및 문서분류가 충분한 것으로 판단된다. 부가하여 신문기사 혹은 문서 간의 표절정도는 <그림 2>에서 제시하고 있는 유사도 곡선에 따라 역산하여 사용해야 하는데, <표 5>에서 서울신문과 동아②의 유사도 50%는 <그림 2>에 따르면 약 80%의 일치도로 환산할 수 있다.

유사도 분석결과표에서 각 문서들은 그룹 내의 다른 문서들과의 유사도가 높은 순으로 정렬한 것으로서, 각 그룹의 좌측 상단에 위치한 문서 혹은 기사는 그룹 내의 다른 문서들과 가장 연관성이 높은 핵심문서(Key Document)가 된다. 시뮬레이션 결과, 세 번째 그룹에서 $tf\text{-}idf$ 방식은 한강을 핵심기사로 판정하였으나, 연관분석방식에서는 연합②를 핵심기사로 판정하였다. 일반적으로 많은 기사들이 연합뉴스를 기반으로 작성한다는 점에서, 연관분석방식이 핵심문서를 보다 적합하게 추천하고 있는 것으로 판단된다.

## Ⅴ. 결 론

인터넷 및 정보통신기술의 발달은 공유정보량의 폭발적인 증가를 가져왔으며, 이는 표절이라고 하는 새로운 사회적 문제를 발생시키고 있다. 이에 따라 문서를 상호 비교하여 문서 간의 표절여부를 검증하는 표절검사 기법에 대한 연구가 활발하게 진행되고 있다. 표절을 검사하는 방법에는 N-gram 방식, 문자열 비교방식, 색인어를 추출하여 가중치를 부여하는 벡터공간모델, 그리고 의미정보를 이용하는 LSA 방식 등이 있다. 대부분의 표절검사에 사용되는 벡터공간모델은 단어의 변형이나 우연한 사용 등에 유연하게 대처하지 못하며, LSA 방식은 한정된 단어를 이용하여 의미적 유사도 측정할 뿐만 아니라 표절하지 않은 문서를 표절로 인식하는 단점이 있다.

이에 따라 본 연구에서는 문장에 사용된 단어 간의 관계를 통해 단어에 의미(semantic)를 부여하며, 우연히 사용된 단어를 표절검사에서 제외하기 위해 데이터마이닝의 연관분석기법을 적용하는 방안을 제안하였다. 형태소분석을 통해 표절검사 대상 문장으로부터 형태소를 추출한 후, 이를 쌍으로 구성하여 문서 간의 유사도를 계산하는 것이다. 몇 가지 사례에 대해 시뮬레이션을 실시한 결과, 표절검사 및 문서분류의 정확도를 크게 향상시키는 것으로 나타났는데, 이것은 형태소 쌍이 각 형태소에 의미를 부가하는 역할을 하기 때문이다.

본 연구에서 제안한 형태소 쌍을 이용한 표절검사 방법이 표절검사의 정확성을 높이는 것으로 나타났지만, 본 연구의 기반이 되고 있는 연관분석방법은 단어를 중심으로 분석하는 $tf\text{-}idf$과 비교할 때 연산에 많은 시간이 소요된다는 단점을 갖고 있다. 따라서 향후의 연구에서는 지수함수의 형태를 갖는 유사도 함수의 특성을 보다 효과적으로 활용하는 방안과 함께 검색엔진 등의 대규모의 문서에 적용 가능하도록 알고리즘의 효율성을 높이기 위한 방안을 연구할 계획이다.

## 참고문헌

김진환, 홍태호, “지식검색 서비스에서 집단지

성 품질이 지속사용 의도에 미치는 영향: 기대일치이론과 신뢰를 중심으로“, 정보시스템연구, 제20권, 제4호, 2011, pp. 1-22.

박선영, 조환규, “성분 정렬을 이용한 한글 유사 문서 탐색 방법”, 한국컴퓨터종합학술대회 논문집, 제38권, 제1호(C), 2011, pp. 228-231.

손윤호, 김인규, 김남규, “연관규칙 마이닝을 활용한 개념적 데이터베이스 설계 자동화 기법”, 정보시스템연구, 제18권, 제4호, 2009, pp. 59～86

신동호, LSA를 이용한 내용기반 검색엔진 시스템, 서울대학교 석사 학위 논문, 2000.

유은지, 김정철, 이춘열, 김남규, “시맨틱 텍스트 마이닝을 위한 온톨로지 활용 방안”, 정보시스템연구, 제21권, 제3호, 2012, pp.137-161.

이동욱, 백서현, 박민지, 박진희, 정혜욱, 이지형, "LSA를 이용한 문장 상호 추천과 문장 성향 분석을 통한 문서요약", *Journal of Korean Institute of Intelligent Systems*, Vol. 22, No. 5, 2012, pp. 656-662.

정석경, 분포정보를 이용한 명사 소프트 클러스터링 연구, 연세대학교 석사학위 논문, 1997.

정영미, 이재윤, “한국어 텍스트내 용어연관성 분석을 위한 기초연구”, 제5회 한국정보관리학회 학술대회 논문집, 1998, pp. 243-246.

조준희, 한국어 문서 표절 검사를 위한 LSA와 N-gram 기반의 유사 문장 판별, 고려대학교 석사학위 논문, 2009.

지정훈, 우균, 조환규, “굼벨 분포 모델을 이용한 표절 프로그램 자동 탐색 및 추적”, 정보처리학회논문지A, 제16-A권, 제6호, 2009, pp. 453-462.

지혜성, 조준희, 임희석, “한국어 문장 표절 유형을 고려한 유사 문장 판별”, 한국컴퓨터교육학회 논문지, 제13권, 제6호, 2010.

황인수, “인터넷 검색과 형태소분석을 이용한 표절검사시스템의 개발에 관한 연구”, 정보기술응용연구, 제16권, 제1호, 2009, pp. 21-36.

Ahmed H. Osmana, Naomie Salima, Mohammed S. Binwahlanc, Rihab Alteebd, Albaraa Abuobiedaa, "An improved plagiarism detection scheme based on semantic role labeling," *Applied Soft Computing*, Vol 12, 1012, pp. 1493-1502.

Agrawal, T., Imielinski T., and Swami A., "Mining Associations between Sets of Items in Massive Databases," *Proceedings of the ACM SIGMOD International conference on Management of Data*, Washington D.C, 1993., pp. 207-216.

Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, Vaclav Snasel, "Survey of Plagiarism Detection Methods," *2011 Fifth Asia Modelling Symposium Conference in Theory and Practice of Digital Libraries*, 2011.

Brijs, T., Swinnen G., Vanhoof K., and Wets G., "Using Association Rules for Product Assortment Decisions: A Case Study," *Proceedings on KDD-99*, ACM, San

Diego, CA, USA, 1999, pp. 254-260.

Brin S. , J. Davis, H. Garcia-Molina, “Copy detection mechanisms for digital documents,” *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, ACM, San Jose, CA, United States, 1995, pp. 398-409.

Donaldson, J. L., Lancaster, A., and P. H. Sposato, “A plagiarism detection system,” *Proceedings of the 20th SIGCSE*. 1981, pp. 21-25.

Fernando Sánchez-Vega, Esaú Villatoro- Tello, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, Paolo Rosso, "Determining and characterizing the reused text for plagiarism detection," *Expert Systems with Applications*, Vol 401, 2013, pp. 1804-1813.

Gabriel Oberreuter, Juan D. Velasquez, “Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style,” *Expert Systems with Applications*, 2013.

Georgina Cosma, Mike Joy, “Evaluating the Performance of LSA for Source-code Plagiarism Detection,” *Informatica*, Vol. 36, 2012, pp. 409-424.

James P. Purdy, "Anxiety and the Archive: Understanding Plagiarism Detection Services as Digital Archives," *Computers and Composition*, Vol 26, 2009, pp. 65-77.

Liu, Y. T., Zhang, H. R., Chen, T. W., & Teng, W. G., “Extending Web Search for Online Plagiarism Detection,” 1-4244-45004/ 07 IEEE, 2007.

Narayanan Shivakumar, “SCAM : A Copy Detection Mechanism for digital Documents,” *Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries*, 1995.

Paul Clough, “Measuring Text Reuse,” *Proceedings of the conference : Association for Computational Linguistics. Meeting*, V.40, 2002, pp. 152-159.

Salha M. Alzahrani, Naomie Salim, and Ajith Abraham, "Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods," *IEEE Transactions On Systems, Man, and Cybernetics –Part C: Applications and Reviews*, Vol. 42, No. 2, 2012.

루씬 한글분석기 오픈소스 프로젝트, http://cafe.naver.com/korlucene

Perez, J. C., http://www.itworld.co.kr

황인수(Insoo Hwang)

저자는 전주대학교 스마트미디어학부 정보시스템전공 교수로 재직하고 있다. 고려대학교 경영학과를 졸업하고 동 대학원에서 경영정보시스템을 전공하여 석사 및 박사학위를 취득하였으며, 산업연구원(KIET) 물류·유통연구센터의 연구원을 역임하였다. 주요 관심분야는 인터넷 정보검색 에이전트, 데이터마이닝, 자연어처리 등이다.

<Abstract>

# A Study on Plagiarism Detection and Document Classification Using Association Analysis

Insoo Hwang

Plagiarism occurs when the content is copied without permission or citation, and the problem of plagiarism has rapidly increased because of the digital era of resources available on the World Wide Web. An important task in plagiarism detection is measuring and determining similar text portions between a given pair of documents. One of the main difficulties of this task is that not all similar text fragments are examples of plagiarism, since thematic coincidences also tend to produce portions of similar text. In order to handle this problem, this paper proposed association analysis in data mining to detect plagiarism. This method is able to detect common actions performed by plagiarists such as word deletion, insertion and transposition, allowing to obtain plausible portions of plagiarized text. Experimental results employing an unsupervised document classification strategy showed that the proposed method outperformed traditionally used approaches.

***Keywords***: Association Analysis, Document Classification, Plagiarism, Plagiarism Detection

* 이 논문은 2013년 12월 11일 접수하여 1차 수정을 거쳐 2014년 9월 19일 게재 확정되었습니다.