한국어 문장 표절 유형을 고려한 유사 문장 판별

지혜성† · 조준희† · 임희석††

요 약

본 논문은 한국어 표절 검사를 위해서 표절의 유형을 분석하여, 유형별 분석 결과를 기반으로 하여 유사 문장 판별 모델을 제안한다. 제안하는 방법은 한국어 문장에 대한 표절 유형 분석 결과를 토대로 LSA와 N-gram을 이용한 유사 문장 검색을 통하여 여러 유형의 표절로부터 견고한 유사 문장 판별 모델을 구현하였다. 제안한 모델의 성능 분석을 위해서 학생들이 인위적으로 작성한 표절 리포트와 표절한 첨부 문서로 실험 데이터를 구축하였다. 성능 비교를 위해서는 기존의 N-gram 모델, 벡터모델, LSA 모델이 사용되었으며, 실험 결과 제안한 모델이 정확률, 재현율, 그리고 F값 척도에서 우수한 성능을 보임을 알 수 있었다.

주제어 : 표절, 유사 문장, N-gram 모델, LSA 모델

A Detection Method of Similar Sentences Considering Plagiarism Patterns of Korean Sentence

Ji Hye-Sung† · Joh Joon-Hee† · Lim Heui-Seok††

ABSTRACT

In this paper, we proposed a method to find out similar sentences from documents to detect plagiarized documents. The proposed model adapts LSA and N-gram techniques to detect every type of Korean plagiarized sentence type. To evaluate the performance of the model, we constructed experimental data using students' essays on the same theme. Students made their essay by intentionally plagiarizing some reference documents. The experimental results showed that our proposed model outperforms the conventional N-gram model, Vector model, LSA model in precision, recall, and F measures.

Keywords : plagiarism, similar sentence, N-gram model, LSA model

† 정 회 원: 고려대학교 컴퓨터교육과
†† 종신회원: 고려대학교 컴퓨터교육과(교신저자)
논문접수: 2010년 08월 13일, 심사완료: 2010년 11월 02일
* 본 연구는 한국연구재단을 통해 교육과학기술부의 뇌과학원천기술개발 사업으로부터 지원받아 수행되었습니다(2010-0029268)

1. 서 론

인터넷이 발전하면서 정보와 다양한 서비스가 빠른 속도로 확산되고 있으며, 많은 사람들이 날마다 인터넷의 방대한 정보를 접하고 있다. 이러한 인터넷의 발전은 유용한 정보와 생활의 편리함을 주었으며 정보의 공유를 통한 새로운 가치 창출과 시간과 공간의 제약을 극복한 정보습득이라는 장점을 주었다. 그러나 많은 정보들은 오히려 정보 과부화(Information overload)를 발생시켜 정보에 대한 검색과 조직화를 어렵게 하였으며[1], 표절(plagiarism)과 저작권 침해라는 새로운 사회적 문제점을 야기 시켰다. 특히 어문저작물의 표절1)은 어린 학생들의 숙제부터 심하게는 학회 논문 및 학위 논문에 이르기 까지 광범위하게 이루어지고 있다. 최근 김병준 전 교육부총리, 마광수 연세대 교수 등의 표절 논란은 표절에 대한 사회적 심각성을 보여주는 대표적인 예로 볼 수 있을 것이다[2]. 그러나 이러한 일이 발생함에도 불구하고 윤리적인 차원에서 표절을 방지하고자 하는 노력이 조금 이루지고 있을 뿐, 표절 검사 및 확인에 대한 연구는 많이 이루어지고 있지 않다.

전부터 표절에 대한 심각성을 미리 알고 표절 방지에 대한 방법을 많이 연구해온 선진국들의 경우는 윤리적으로 많은 교육을 실시하고 있으며, 표절 검사 시스템들을 개발하여 표절에 대해 엄중하게 대처하고 있다. 우리나라 역시 표절위원회를 발족하는 등 표절에 관련된 문제점을 인식하고 표절방지를 위한 노력을 시작하였으나[3], 한국어의 특성에 맞추어 적합한 표절 검사에 대한 연구는 많지 않은 실정이다.

정보 공유를 통한 장점을 승화시키고 표절과 같은 문제점을 미리 차단하기 위해서는 표절 검사 시스템 구축과 같은 대책을 마련하는 것이 양질의 정보를 공유하고 독창적인 연구가 이루어지는데 도움이 될 것이다. <표 1>은 표절 판정에 대한 기준을 나타낸 것이다[1].1)

<표 1> 표절 판정 기준

표절 판정 기준
1.다른 사람의 글을 인용 표시 없이 그대로 사용할 경우
2.표, 그림, 모식도 등의 자료를 인용 없이 사용할 경우
3.다른 사람의 실험보고서를 참고하여 작성하는 경우
4.같은 종류의 자류를 두 개 이상의 학술지에 투고하는 경우
5.한번 발표한 논문을 다른 언어로 다시 작성하는 경우
6.다른 사람의 정보를 자신의 것처럼 위장하여 사용할 경우

<표 1>과 같은 판정기준이 있더라도 표절을 찾아내는 것은 쉬운 일이 아니다. 사람이 모든 문서에 대한 정보를 찾아서 문서의 표절 여부를 판단할 수 있지만, 수많은 정보에서 작성 문서에 대한 정보를 찾기는 쉽지 않으며 많은 양의 정보를 사람이 수동으로 검사하기는 사실상 불가능에 가깝다[15]. 따라서 이러한 단점을 개선하기 위해서는 자동으로 문서의 유사도를 검색하여 표절 여부를 판단할 수 있는 시스템의 개발을 필요로 한다. 표절 검사를 위한 유사도 검색 시스템은 일반적인 정보검색 시스템과는 달리 높은 정확도와 재현율을 가지는 검색 방법을 필요로 하며, 어순 변경이나 단어 치환과 같은 많은 표절의 유형들을 분석하여 그로부터 견고한 시스템을 구축하여야 한다.

기존에 많이 개발되어 온 표절 검사 방법은 주로 영어 등 외국어에 맞추어진 것으로 한국어 특성에 맞추어 개발한 검사 방법이 아니다. 또한 표절은 어순 변경이나 구조 변경 혹은 단어 치환 등과 같은 여러 가지 방법을 사용하여 문장을 의미적으로 유사하게 변형시켜서 검사 시스템에 발각되지 않도록 지능적으로 이루어지고 있다. 최근 표절은 원문 전체를 그대로 복사하는 방법으로 사용하는 것이 아니라 문장 단위로 필요한 부분만을 사용하거나 그 문장을 구성하는 단어를 의미적으로 유사한 단어로 바꾸어서 사용하는 것과 같이 문장을 변형하여 지능적으로 표절로부터 벗어나고자 하는 시도가 이루어지고 있다. 따라서 표절 검사는 문장 단위로 의미적 유사도를 고려한 검사를 하는 것이 보다 정확한 표절 검사가 될 것이다.

본 논문은 한국어 문장 표절 유형을 분석하고 분석된 모든 문장 표절 유형을 찾을 수 있는 견고한 한국어 유사 문장 판별 모델을 제안한다.

1) 이후부터 특별한 언급을 하지 않는 한 '표절'은 어문저작물의 표절을 의미함

2. 관련 연구

2.1 기존 유사 문장 판별 방법

2.1.1 N-gram 방식

N-gram이란, 인접한 N개의 음절을 말하며, N-gram 방식은 두 문장 내에 존재하는 N-gram을 추출하고 그것들 중에서 얼마나 많은 N-gram이 일치하느냐에 따라서 문장의 유사 여부를 판단하는 방법이다[13]. 다음은 N-gram 추출 방법에 대하여 설명한다. 문서에 있는 문장을 먼저 빈칸, 마침표, 쉼표 등을 구분자로 하여 모든 어절을 추출하고, 추출한 어절들에 대해 N-gram을 추출한다. 예를 들면 “표절검색”이라는 단어의 bi-gram은 “표절”,“절검”,“검색”이며, tri-gram은 “표절검”,“절검색”이다. 어절의 음절 수가 N보다 큰 경우에는 여러 개의 N-gram으로 분리되고, 작은 경우에는 하나의 N-gram으로 취한다. 따라서 문서에서 철자 오류가 있더라도 문장은 유사한 것으로 검색될 가능성이 높다. 그러나 이 방식은 문장의 양이 많아질수록 많은 수의 N-gram이 생성되기 때문에 많은 저장 공간이 필요하며, 전혀 다른 문장임에도 불구하고 생성된 N-gram에 의해서 일치하게 되는 부적합 현상이 생길 수 있다.

2.1.2 문자열 비교 방식

문자열 비교를 이용한 연구는 기존에 영어권에서 만들어진 시스템에서 널리 사용되고 있는 방법으로서, 비교하고자 하는 두 문장과 문장 간에 정확한 문자열 비교가 아닌 문장의 변형을 고려하여 두 문장 간에 차이를 허용하는 문자열 비교를 통해서 문장과 문장 간을 비교하고 두 문장 사이에서 일치하는 결과가 문장 내에 포함되는 정도를 가지고 문장의 유사여부를 판단한다[13]. 문자열 비교는 기존의 많은 시스템에서 사용되는 것에서 알 수 있듯이, 어순의 변경이나 단어의 삽입이 적게 이루어진 경우에서는 높은 신뢰도를 가지지만, 영어에 비하여 어순이 자유로운 한국어에 대해서는 그 어순 변경을 통한 변형의 경우 견고하지 못한 면이 있다. 또한, 문장 검사할 때 최소 매칭 길이의 결정에 있어서, 최소 매칭 길이를 작게 설정해 주면 문장의 변형이 많이 이루어진 경우에도 문장의 유사여부를 판단할 수 있으나, 그 판단의 확실성은 떨어지게 된다. 반면에 최소 매칭 길이를 크게 설정해 주었어도 문장의 두 문장의 유사여부 판정의 확실성은 보장할 수 있으나 , 문장의 변형에 견고하지 못한 단점이 있다. 또한 효율성 면에서 문장이 n개가 존재하는 경우 약 n*(n-1)번 문자열 비교를 수행해야 하므로 문서 집합이 증가하는 경우 시간이 오래 걸리는 단점을 가지고 있다.

2.1.3 벡터 공간 모델 방식

이 방식은 정보검색 모델의 한 종류인 벡터 공간 모델을 응용하여 문장의 유사도 값을 계산하여 유사 여부를 판단하는 방법이다[14]. 각 문장을 이루는 색인어를 추출하여 벡터 공간 상의 벡터로 표현하여 문장과 문장 사이의 유사도를 계산한다. 만약 문장 내에 색인어가 포함되어 있으면, 벡터 내의 해당차원은 0이 아닌 가중치의 값을 가지게 된다. 가중치 값은 많은 방법이 있으나 주로 가장 잘 알려진 TF-IDF 가중치 방법이 주로 많이 사용된다. 가중치로 인하여 표현된 문서를 N차원의 벡터 값으로 나타낸 후, 문장과 문장 간의 유사도를 계산한다. 유사도 계산은 다이스 유사계수, 쟈카드 유사계수, 내적 계수, 코사인 유사계수 등이 이용되는데 주로 코사인 유사 계수가 가장 많이 사용된다. 벡터 공간 모델은 계산이 비교적 간단하며 정규화 된 유사도 값을 얻을 수 있다는 장점이 있지만 색인어로 추출되는 키워드가 정확히 일치되어야 하며, 비슷한 의미를 지니고 있더라도 그것을 유추해 낼 수 없다는 단점이 있다.

2.1.4 LSA(Latent Semantic Analysis) 방식

이 방식은 문장에 단어나 어절을 유사한 의미로 변형을 가한 경우에 검색되지 않는 벡터 공간 모델의 단점을 극복하기 위해 응용되는 방법으로 개념적으로 공기정보를 이용하여 단어들의 의미 관계를 찾아 문장의 유사여부를 판단한다. 즉, 단어들의 국소 문맥 정보를 이용해서 단어들 간의 의미 관계를 찾아 낼 수 있다[6]. LSA는 이론적으로 선형대수학의 SVD 라는 통계적 기법과 차원 감소가 사용되며, 이를 통하여 각 단어나 문장

이 고차원의 의미 공간에 하나의 벡터로 표현된다. 나타난 벡터간의 유사도는 벡터 공간 모델과 유사한 방식으로 계산되며 주로 코사인 유사 계수가 많이 이용된다. LSA는 의미적 관계를 찾지 못하는 벡터공간 모델의 단점을 극복할 수 있다는 장점을 가지지만 형태소나 어절 수준의 정보를 고려하는 N-gram이나 벡터 공간 모델과 비교하여 다른 문장임에도 불구하고 의미적 유사성에 의하여 유사 문장으로 검색되는 것과 같이 정확도가 떨어지는 단점이 있다.

3. 한국어 문장의 표절 유형

문서와 문서간의 표절은 원문의 문서 내용을 그대로 가져다 사용하는 것에서 점차 교묘하게 변질되었으며, 좀 더 지능적이고, 교묘한 방법으로 표절 의혹으로부터 벗어나고자 하는 방향으로 발전하였다[7]. 최근에는 각 문서에서 좋은 내용의 문장을 추출하여 사용하는 이른바 짜깁기 형식의 표절 형태가 주로 보이고 있다. 표절을 하는 방식과 유형은 나날이 지능적으로 변하고 있으며, 표절을 근절하기 위해서는 기존의 표절 문서들에 대한 연구가 필요하다. 따라서 여러 유형의 표절 방법들을 모두 감지하기 위해서는 표절 문서 내에서 보이는 문장들의 표절 유형에 관한 연구가 선행되어야 한다.

<표 2> 유형별 표절 예제

표절 유형	표절 예제 문장
단어 치환	지난 6.10 범국민대회 참가자들에게 방패를 휘두른 의경 2명에 대해 경찰이 징계 절차에 착수했습니다.
	지난 6.10 범국민대회 참여자들에게 방패로 공격한 의경 2명에 대해 경찰이 징계 절차를 시작했습니다.
어순 변경	음주와 비만, 운동부족 등도 장암의 원인으로 거론되고 있지만, 적색육 및 가공육의 섭취가 특히 위험한 발병 요인이라는 과학적 증거가 최근 부쩍 많이 제시되고 있다.
	적색육 및 가공육의 섭취가 비만, 운동부족등이 원인이 되는 장암의 발병 요인이라는 과학적 증거가 최근 부쩍 많이 제시되고 있다.
원문 요약	계산을 하거나 하지 않는 것도 필요하지만 중요한 사실은 외부의 명령어를 읽어 그것을 수행하는 하나의 완결적인 제어 메커니즘을 갖게 되었다는 점이다
	계산도 필요하지만 중요한 사실은 명령어를 읽어 수행하는 제어 메커니즘을 갖게 되었다는 점이다

본 논문에서는 보다 정확한 표절 문서 검사를 위해서 표절 문서 내에서 보이는 문장들의 표절 유형에 대하여 연구 하였다. <표 2>는 한국어 표절 문서에서 발견되는 문장 유형을 나타낸 것으로 원문복사, 단어치환, 어순변경, 문장요약 유형으로 분류하였으며, 유형별 특징은 다음과 같다.

3.1 원문 복사 유형

원문 복사 유형은 원문에서 존재하는 문장을 그대로 복제(copy)하여 표절 문서에 적용한 유형을 의미한다. 주로 표절에 익숙하지 못한 사람들이나 학생들의 숙제나 리포트에서 주로 나타나며, 출처를 명시하고 원문을 그대로 사용한 경우도 포함된다. 또한, 새로운 연구나 논문에 자신의 이전 연구 내용을 다시 재사용하는 경우를 자가 표절이라고 하며 이 경우에도 원문 복사 유형에 포함된다.

3.2 단어 치환 유형

단어 치환 유형은 문장의 전체적인 의미를 훼손하지 않는 범위 안에서 어절 혹은 어절 내의 형태소나 명사류의 단어를 추가하거나 비슷한 의미의 다른 단어로 바꾸는 유형을 말한다. 즉, 문장의 핵심적 의미를 가지는 단어를 같은 의미를 지니는 다른 단어로 치환하는 것으로, 가장 많이 사용되고 있는 표절 유형 중 하나이다.

3.3 어순 변경 유형

어순 변경 유형 역시 문장의 전체적인 의미나 흐름을 그대로 반영하면서 문장 내의 어절 또는 구(phrase)의 위치를 변경하여 다른 문장인 것처럼 위장하는 표절 유형이다. 이러한 어순 변경 유형의 경우도 문장을 만들 때 저자의 독창적인 아이디어가 적용되었다고 볼 수 없기 때문에 표절 문장으로 구분한다.

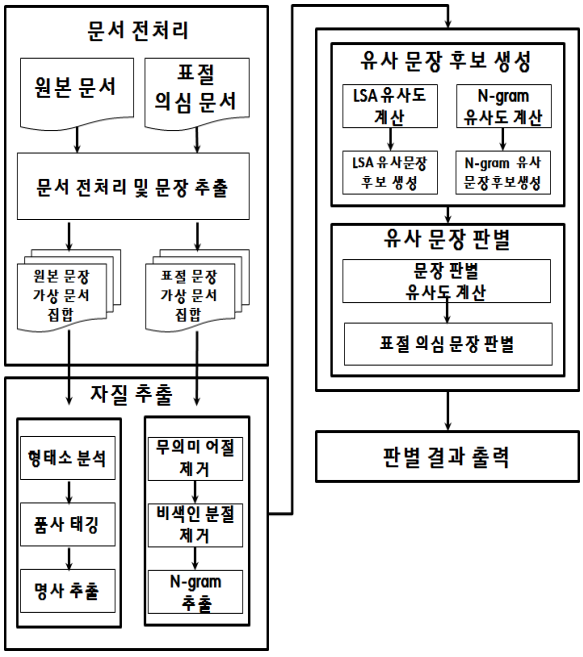
3.4 문장 요약 유형

문장 요약 유형은 글의 전체적인 의미를 손상시키지 않는 범위 내에서 특정 어절을 삭제하거나 변형을 통하여 다른 문장처럼 보이도록 하는

표절 유형을 말한다. 문장의 의미를 이루는 핵심적인 단어를 제외한 부분을 삭제하기 때문에 원문과 비교하여 유사도가 높게 나오지 않는다.

4. 유사 문장 판별 모델

본 논문은 문장 간의 의미적 유사도를 검색하여 유사한 문장을 추출하고, 동시에 형태적 유사도를 검색하여 문장을 추출함으로서 문장의 의미적 유사여부와 형태적 유사여부를 동시에 분석하여 보다 표절에 견고한 시스템 구현을 목적으로 한다. <표 2>에서 제시한 것처럼 표절 유형에는 원문 복사, 단어 치환, 어순 변경, 문장 요약과 같은 유형이 존재한다. 원문 복사를 제외한 유형에는 기존의 검사 방법들이 모두 취약한 점을 보이고 있다. N-gram과 문자열 비교 방식의 경우에는 형태적으로 문장에 변형을 많이 가하는 단어 치환과 문장 요약의 경우에 취약한 면을 보였으며, 벡터 공간 모델 방식은 자질 선정 방식에 따라 다르나 주로 단어 치환 유형과 같이 의미적으로 유사하게 문장에 변형을 가한 경우에 성능이 떨어지는 단점이 있다. LSA 분석 기법을 이용한 방식은 의미적 유사성을 이용한 변형에도 견고한 면을 보였으나, 그 반대로 의미적으로 유사할 뿐, 실질적인 표절이 아님에도 불구하고 유사 문장으로 인식하는 단점이 나타났다. 본 논문에서는 기존 방식의 단점을 극복하고 여러 가지 표절의 유형에도 견고한 유사 문장 판별 모델을 제안하고자 한다. 제안하는 모델은 단어 치환과 같이 의미적 유사성을 이용한 표절 유형을 극복하기 위해서 LSA 분석 기법을 이용함과 동시에 어순 변경과 같은 형태적 변형에 우수한 성능을 보이는 N-gram 비교 방식을 이용하여 여러 표절 유형으로부터 견고한 문장 판별 모델을 구축하였다. <그림 1>는 본 논문에서 제안하는 유사 문장 판별 모델의 구성을 나타낸 것이다.



<그림 1> 유사 문장 판별 모델 구성도

4.1 문서 전처리

문서 안에서 표절은 문서를 전체적으로 표절하는 것이 아니라 문서 안에서 필요한 특정 부분. 즉, 문장 단위로 표절이 이루어지고 있다. 문장 단위로 이루어지는 표절을 찾아내기 위해서는 문서를 전체적으로 비교하는 것이 아니라 문장 단위로 나누어서 유사도를 계산하여 일정 이상의 유사도를 가지는 문장들을 표절 문장으로 추출하여야 한다. 이 방식은 문장별로 나누어져 검사하기 때문에 여러 개의 문서에서 표절이 이루어진 경우에도 어느 부분에서 표절을 했는지에 대한 여부를 쉽게 판단할 수 있다는 장점이 있다[7]. 본 논문은 문장 구분을 위하여 원본 문서와 표절이 의심되는 문서를 마침표(.)등의 문장부호를 통하여 문장으로 나눈다. 그러나 문장의 길이가 너무 짧은 경우에는 문장 내에서 의미를 내포하고 있다고 보기 어렵기 때문에 원문과 6개 단어 이상 연속 동일할 경우를 기준으로 하여 6어절 이상의 문장을 추출한다. 추출된 문장들은 각각 하나의 문서로 간주하여 각각 원본 문서 O_i에 대한 가상 문서 집합 $O_i = \{od_{i1}, od_{i2}, \ldots, od_{in}\}$으로 n개의 문서 집합이 만들어지며 표절 문서 P_j에 대한 가상 문서 집합 $P_j = \{pd_{j1}, pd_{j2}, \ldots, pd_{jm}\}$으로 m개의 문서 집합이 만들어진다. 따라서 총 m+n 개의 가

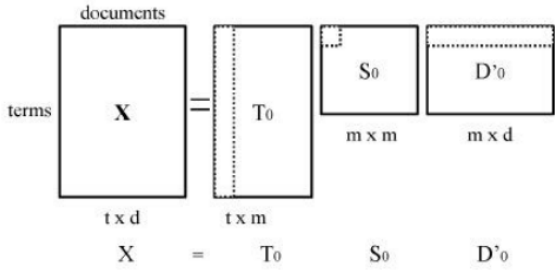
상 문서가 만들어지게 되며, 이 가상 문서들에서 자질 추출을 하도록 한다.

4.2 자질 추출

문장의 유사도 계산을 위해서는 자질 추출을 필요로 한다. 본 연구에서는 LSA를 이용한 유사도 계산과 N-gram을 이용한 유사도 계산 두 가지 방법의 계산이 이루어지기 때문에 각각 방식에 필요한 자질을 추출하여야 한다. LSA를 이용한 유사도 계산에서는 문장을 나타낼 수 있는 자질들 중에서 주로 의미를 나타내는 명사와 고유명사를 자질로 추출 하였으며, N-gram 유사도 계산에서는 문장을 어절로 분리하고, 불용어 리스트를 이용하여 자질로서 무의미한 어절들을 삭제한 후, 나머지 음절로부터 음절 bi-gram과 tri-gram을 추출한다.

4.3 LSA를 이용한 유사 문장 후보 생성

본 연구는 문장의 의미적 유사 여부를 판단하기 위하여 문서에서 문장별로 분리 후, LSA를 이용하여 만들어진 의미공간에서 개념들 간의 유사도를 측정하였다. 자질 추출을 통하여 추출한 명사를 색인어로 하여 유사도 계산을 위한 가중치를 계산하였으며, 가중치 계산식은 TF*IDF 값을 사용하였다. 가중치 계산 후에 벡터로 표현된 가상 문서들을 단어-문서 행렬로 나타낸 후, LSA를 이용하여 기존에서는 나타나지 않는 의미적 유사성을 찾도록 하였다. 단어-문서 행렬은 SVD에 의하여 <그림 2>와 같이 3개의 행렬로 분해된다.



<그림 2> 단어-문서 행렬 SVD

t는 단어의 개수이며, d는 문서의 개수, m은 행렬 X의 rank이다. 여기서 k(k<m)개만큼의 벡터만을 사용하여 차원축소를 한다. 위의 결과를 통해 단어와 단어, 문서와 문서, 단어와 문서의 관계를 비교할 수 있다[6]. 본 논문에서는 문장의 유사도를 비교하기 위해서 추출한 문장들을 각각 문서로 가상하고, S와D 행렬의 곱으로 생성되는 문서-문서 행렬을 이용하여 문장과 문장 사이의 유사도를 계산하였다. 또한, k 값에 따라 행렬을 구성하는 값이 변하므로 실험을 통하여 k값을 결정하였다.

LSA에 의하여 구해진 문서-문서 행렬을 유사도 계산식을 통하여 원본 문서의 문장과 일정 수준 이상의 유사도를 가지는 표절 의심 문서의 문장을 추출한다. 유사도 계산은 코사인 유사도 계산식을 사용하였으며, 유사도 계산을 마친 가상 문서들은 유사 가상 문서 추출 함수를 통하여 표절 여부를 판별하도록 한다. 유사 가상 문서 추출 함수는 $\Phi_{LSA}(O_i, P_j)$로 다음과 같이 정의하였다.

$$\phi_{LSA}(O_i, P_j) \overset{\text{def}}{\Rightarrow} \{(\langle od_{im}, pd_{jn}\rangle, SIM_{LSA}(od_{im}, pd_{jn}))|SIM_{LSA}(od_{im}, pd_{jn}) \geq \delta_1\} \quad (1)$$

식(1)은 LSA에서 계산되어진 일정 정도(δ_1) 이상의 유사도 값을 가지는 가상 문서를 표절 의심 문서로 추정하는 식이며, 여기서 추출된 문서들은 표절 의심 문서로 추정하고, 다음 단계에서 유사 문장 판별 계산을 통하여 문장에 대한 표절 여부를 판별하게 된다.

4.4 N-gram을 이용한 유사 문장 후보 생성

본 연구에서는 LSA 유사도 계산과는 별도로 문장에 대한 N-gram 중복 검사를 실시하여 보다 정확한 표절검사를 하고자 하였다. 검사는 음절 tri-gram 과 bi-gram을 가지고 하였으며, 각 문장별 N-gram 유사도 계산 값에 대한 계산식 $SIM_{Ngram}(od_{im}, pd_{jn})$은 식(2)와 같다.

$$SIM_{Ngram}(od_{im}, pd_{jn}) = \alpha \cdot \left(\frac{2a}{M+N}\right) + \beta \cdot \left(\frac{2b}{V+W}\right) \quad (2)$$

a : 중복되는 tri-gram 수
b : 중복되는 bi-gram 수
M : od_{im}의 총 tri-gram 수
N : pd_{jn}의 총 tri-gram 수

V : od_{im}의 총 bi-gram 수
W : pd_{jn}의 총 bi-gram 수
α, β : N-gram 가중치 (단, $\alpha + \beta = 1$)

α와 β의 값은 각자 tri-gram과 bi-gram의 가중치 정도를 나타내며 두 가지 중에서 어디에 더 가중치를 부여하는가에 대한 값들이다. 본 연구에서는 α와 β 값의 최적 값을 구하기 위해 실험하였으며 그 결과 α 값은 0.6, β 값은 0.4 일 때 가장 좋은 성능을 보임을 확인할 수 있었다.

식(2)의 결과 값이 일정 정도(δ_2) 이상의 유사도 값을 가지는 가상 문서를 표절 의심 문장으로 추정하는 함수는 다음과 같다.

$$\Phi_{Ngram}(O_i, P_j) \overset{\text{def}}{\Rightarrow} \left\{ \begin{matrix} (\langle od_{im}, pd_{jn} \rangle, SIM_{Ngram}(od_{im}, pd_{jn})) \\ | SIM_{Ngram}(od_{im}, pd_{jn}) \geq \delta_2 \end{matrix} \right\} \quad (3)$$

여기서 유사 가상 문서로 추정된 가상 문서들은 다음 단계에서 유사 문장 판별 계산을 통하여 문장에 대한 표절 여부를 판별하게 된다.

4.5 유사 문장 판별

앞에서 유사도 계산에 따라 서로 다른 가상 문서들이 표절 의심 문서로 검색되었다. 원본 문서 O_i에 대한 LSA와 N-gram 유사도 계산에 의하여 추출된 가상 문서의 집합을 가지게 되며 각각의 값은 LSA 유사도 계산과 N-gram 유사도 계산에서 추출된 원본 문장과 유사 문자의 쌍으로 이루어져 있다. 이 두 가지의 서로 다른 결과 값을 기반으로 표절 의심 문장을 판별하는 함수는 다음과 같다.

$$\Phi_{result}(O_i, P_j) \overset{\text{def}}{\Rightarrow} \left\{ \begin{matrix} < od_{im}, pd_{jn} > | (< od_{im}, pd_{jn} > \in \Phi_{LSA}(O_i, P_j) \vee \\ < od_{im}, pd_{jn} > \in \Phi_{Ngram}(O_i, P_j)) \\ \wedge \alpha \cdot SIM_{LSA}(od_{im}, pd_{jn}) + \beta \cdot SIM_{Ngram}(od_{im}, pd_{jn}) > \delta_3 \end{matrix} \right\} \quad (4)$$

LSA 유사도 계산과 N-gram 유사도 계산에서 계산된 유사도를 가지고 최종 결과를 추출하는 위 식의 α와 β 값은 각각 LSA와 N-gram에 부여하는 가중치 값을 나타내며, 값들이 높을수록 그 값에 대한 신뢰가 더 높아지는 것을 의미한다. 즉, 식(1)에서 추출된 표절 의심 문장과 식(3)에서 추출된 표절 의심 문장이 일치하고, 두 유사도 값이 일정 정도(δ^3)를 넘으면 문장은 표절된 것으로 판단한다. 이 계산을 통해 두 문장이 완전 동일하면 1의 값을 가지며 관계가 적을수록 값이 줄어들게 되며, 완전히 상관이 없는 경우, 즉 일치하는 부분이 한 군데도 없는 경우에는 0의 값을 가지게 된다.

5. 실험 및 평가

5.1 실험 데이터

실험은 수업에서 학생들로부터 임의적으로 작성된 표절 문서 집합과 표절에 사용된 원본 문서를 대상으로 하였으며, 그 구성은 <표 3>과 같다.

<표 3> 실험 데이터

표절 문서 수	50개
평균 문장 수	184개
참조 문헌 수	10개

표절 문서는 '폰 노이만'이라는 주제를 가지고 50명의 학생들이 미리 지정된 참조 문헌 10개를 이용하여 인위적인 표절을 통하여 작성되었다. 문서 작성 전에 표절 유형에 대한 사전 교육을 통하여 4가지 표절 유형에 맞추어 지능적인 표절이 되도록 유도하였다. 표절된 문서의 수는 총 50개의 문서이며 평균 184개의 문장으로 구성되어 있다.

본 연구에서 평가 방법은 문장 유사도 검색이 수행된 문장에 대한 정확도(P)와 재현율(R) 그리고 조화 평균(F-measure) 값을 사용하였으며, 각각의 계산식들은 다음과 같다.

$$\text{정확률}(P) = \frac{\text{시스템이 찾아낸 정답 개수}}{\text{시스템이 찾아낸 유사문장 판별 개수}}$$

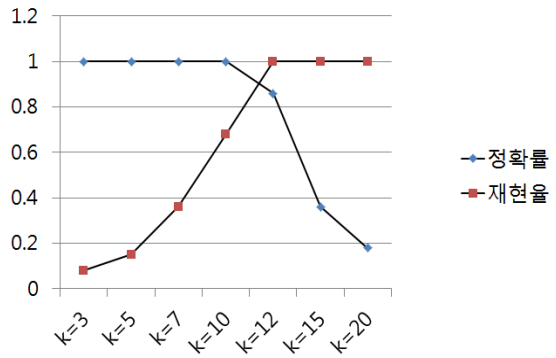
$$\text{재현율}(R) = \frac{\text{정답 집합 개수}}{\text{시스템이 찾아낸 정답 개수}}$$

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}}$$

F 함수의 값은 0~1 사이의 값을 가지며, 이 값은 전혀 연관성이 없을 경우에는 0이 되며, 완전 일치할 경우 1이 된다. 또한, 재현율과 정확률이 모두 높아야 F 값도 높아진다. 따라서 F 값을 최대로 하는 것은 재현율과 정확률 사이의 가장 적절한 값을 구하는 방법으로 볼 수 있다.

5.2 LSA k값 결정 실험

LSA에서 차원축소를 위한 k값은 문서의 수와 문서를 이루는 단어의 수에 따라 적절한 값이 달라지기 때문에 실험을 통하여 적절한 값을 선택하여야 한다. 본 논문에서는 k 값의 변화에 따른 정확률과 재현율을 통하여 k 값을 결정하였으며 결과는 <그림 3>과 같다.



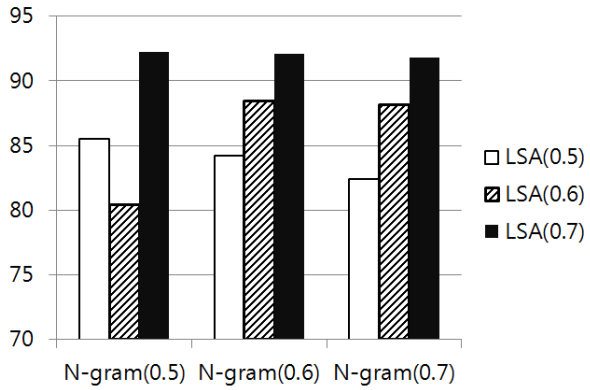
<그림 3> k값 결정 실험 결과

실험 결과 k값이 작아질수록 재현율이 떨어지며 커질수록 정확률이 떨어지는 것을 볼 수 있다. 그 이유는 사용된 k값이 작을수록, 즉 LSA에서 고려하는 개념 공간의 차원이 작을수록 의미적으로 유사한 단어들의 차원이 많이 합쳐져서 정답에 비하여 시스템이 찾아낸 정답의 수가 많기 때문에 정확률은 높더라도 재현율이 낮아지는 것으로 보이며, 반대로 k 값이 커질수록 차원이 적게 합쳐지기 때문에 재현율은 높아지지만 상대적으로 정확률이 낮아지는 것으로 보인다. 실험결과 정확률과 재현율이 가장 이상적으로 나타난 부분은 k=12이지만 본 논문에서는 표절 검사를 위한 정확률에 더 초점을 두기 때문에 k=10으로 설정하였다. 그러나 이것은 실험 데이터에 따라 다르기 때문에 모든 상황에서의 k값이라고 하기 보다는 "폰노이만 이라는 주제를 가진 표절 문서"라는 상황에 맞는 k값으로 보아야 한다.

5.3 표절 검사 실험

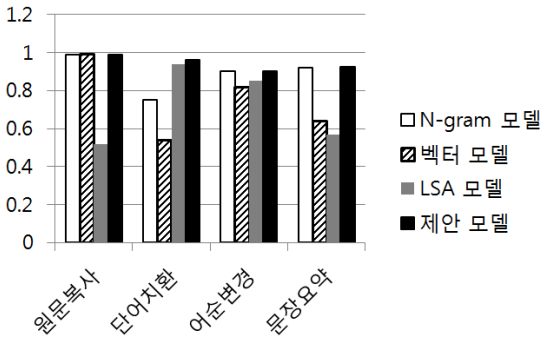
본 연구에서는 기존 비교 방식을 사용했을 때보다 얼마나 성능이 향상되는가에 대해 알아보기 위하여 문자열 비교 방식을 제외한 N-gram, 벡터 모델, LSA, 그리고 제안하는 시스템에 대한 정확률 및 재현율 그리고 조화평균을 분석하여 성능 평가를 실시하였다. 문자열 비교 방식은 한국어의 표절 검색에는 성능이 낮기 때문에 실험에서 제외하였다[5]. 또한 문장 유형별로도 정확률 및 재현율, 조화평균을 분석하여 유형별로 시스템에 대한 성능을 분석하였다.

<그림 4>는 LSA 유사도 임계 값(δ_1) 및 N-gram 유사도 임계 값(δ_2)의 변화에 따른 F-measure 값의 변화를 나타낸 것으로 LSA의 임계 값을 0.7, N-gram의 임계 값을 0.5로 설정하였을 때 가장 좋은 결과를 보여주었다. 이 이유는 LSA의 임계 값이 작아지면 문장과 문장이 의미적으로 유사하지만 실질적으로 표절이 아님에도 불구하고 유사 문장으로 추출하기 때문으로 보여진다. 또한 N-gram의 임계 값이 클수록 형태적 유사성만을 확인하므로 어순 변경 및 문장 요약 유형에 대한 검사 결과를 반영하지 못하여 나타나는 현상으로 보인다.



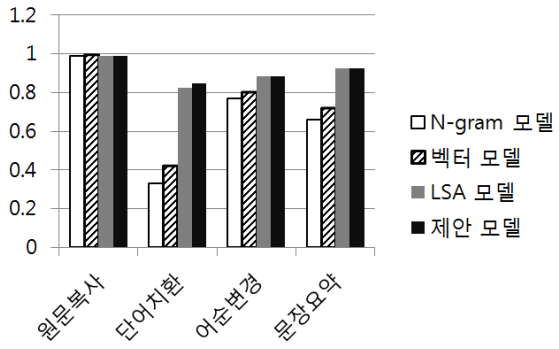
<그림 4> 임계 값 변화에 따른 F-measure 값

위의 실험 결과를 반영하여 LSA의 임계 값을 0.7, N-gram의 임계 값을 0.5로 설정하여 본 연구에서 제안한 문장 판별 모델로 표절 문서를 검사하였다. 본 연구에서는 N-gram 비교 방식, 벡터 모델 방식, LSA 방식을 이용한 시스템과 본 연구에서 제안하는 모델을 표절 유형별로 시스템의 정확률과 재현율을 비교하여 제안하는 모델의 장점을 나타내고자 하였으며 결과는 다음과 같다.



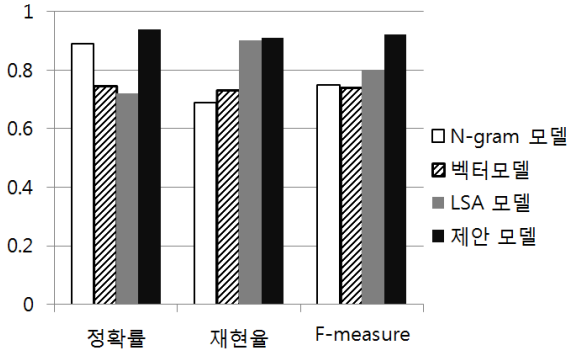
<그림 5> 유형별 정확률 비교

<그림 5>는 표절 유형에 따른 각 모델별 정확률을 비교한 것이다. 원문복사 유형과 문장요약 유형의 경우 N-gram 모델과 벡터 모델에서는 모두 찾아낼 수 있었지만, LSA 모델에서는 그 절반 정도의 정확률을 보여주었다. 그 이유는 LSA 모델은 원문 복사를 한 문장 외에도 의미적으로 유사한 다른 문장들까지 모두 검색하기 때문에 상대적으로 재현율은 높지만 정확률 면에서 떨어지기 때문이다. 또한 단어치환 유형의 경우에는 형태적 검사를 주로 하는 N-gram 모델과 벡터 모델이 LSA 모델보다 현저히 떨어지는 결과를 보여주고 있다. 본 연구에서 제안한 모델은 판별식을 통해 유사도를 계산함으로서 기존의 방식보다 유형별로 높은 정확률을 나타내는 것을 볼 수 있다.



<그림 6> 유형별 재현율 비교

<그림 6>은 표절 유형에 따른 각 모델별 재현율을 비교한 것이다. 단어치환 유형의 경우 N-gram 모델과 벡터모델은 형태적으로 치환된 단어에 의하여 유사 문장을 찾아내지 못하였기 때문에 재현율이 매우 낮게 나타는 것을 볼 수 있다. 그러나 LSA와 본 연구에서 제안한 모델의 경우는 LSA를 이용한 공기정보에 따른 의미적 검사방법을 통하여 단어치환의 유형에도 좋은 성능을 보여주었다.



<그림 7> 모델별 성능 비교

<그림 7>은 각 모델의 정확률과 재현율, F-measure 값을 비교한 것이다. N-gram 모델의 경우에는 전체적으로 재현율이 떨어지는 단점으로 인하여 F-measure 값이 평균 0.75 값을 얻었으며, 벡터모델의 경우에는 정확률과 재현율 두 값이 비슷하게 나타났으나 평균 0.73의 값으로 비교 모델 중 가장 낮은 값을 보였다. 그리고 LSA 모델의 경우 정확률이 떨어지는 단점으로 인하여 평균 0.80 값을 얻었다. 본 논문에서 제안하는 모델은 여러 유형으로부터 견고하게 작동하는 것을 볼 수 있었으며, 평균 0.92의 F-measure 값을 얻

을 수 있었다. 기존의 모델도 대부분 완벽하게 찾아낸 원문 복사 유형을 제외할 경우, 기존 모델의 F-measure 값은 더 떨어지는데 반해, 제안하는 모델의 경우에는 차이가 없는 것으로 나타났다. 이 이유는 원문 복사 유형의 경우는 원문과 표절 문장의 변형이 없으므로 대부분의 모델에서 유사도 검사가 용이하여 높은 재현율과 정확률을 나타내기 때문이다.

본 연구에서 제안하는 문장 판별 모델은 기존의 N-gram 모델, 벡터모델, LSA 모델에 비하여 표절 유형별로 모두 높은 정확률과 재현율을 보여주고 있으며, 이는 기존의 하나의 방식만을 사용하였을 때보다 성능이 개선됨을 알 수 있었다.

6. 결론 및 향후 과제

최근 많은 표절 및 유사 행위로 인하여 사회적 혹은 도덕적으로 문제가 야기되고 있다. 그러나 표절검사가 제대로 이루어지지 않음에 따라 창조적인 연구가 진행되지 않고, 간단하게는 학생들의 숙제부터 심각하게는 논문에 이르기까지 표절이 이루어지고 있다는 점은 사회적으로 심각한 문제이며 꼭 해결해야만 하는 문제라고 볼 수 있다. 이러한 문제점을 해결하기 위한 사항으로 표절 예방을 위한 교육과 표절 여부를 판단하는 시스템 구축 등을 들 수 있다.

본 논문에서는 한국어 표절 검사를 위한 문장 유사도 검색에 있어서 효과적인 검사를 위해서 문장의 표절 유형을 분류하였고 , 그 유형별로 유사도를 검사하여 표절 문장을 판별하는 유사 문장 판별 모델을 제안하였다. 본 논문에서 제안하는 모델은 표절 문서를 효과적으로 비교하기 위해서 문장별로 비교하는 방법을 선택하였으며, 문장의 형태소 및 어절 변형 및 단어 치환에 견고하게 작동하고 문장 유사도 검색을 효율적으로 수행하기 위해 LSA와 N-gram을 이용한 문장 유사도 검색을 하였다. 거기서 발생할 수 있는 정확률 및 재현율의 문제점은 본 연구에서 제안하는 모델을 통해서 해결 할 수 있음을 보였고, 실험을 통해 재현율과 정확률이 향상되는 것을 알 수 있었다. 또한 N-gram만을 이용하여 검사하였을 때 판별할 수 없었던 단어 치환이나 문장 요약 등의 형태에 대하여서도 LSA를 이용한 문장 유사도를 통하여 해결할 수 있는 것을 보였으며, LSA에서 발생하는 정확률 문제는 N-gram을 통하여 보안할 수 있음을 보였다.

향후 앞에서 말한 4가지 유형의 표절 유형뿐만 아니라 다른 유형의 표절 유형에도 견고히 작동하는 문장 검색 시스템이 개발되어야 할 것이며, [4]와 같이 다른 표절 검색 방법과의 비교 연구를 통해 개선해나가야 할 것이다. 또한, 문장 표절 검색뿐만 아니라 예제 기반 MT, 영작문 도우미 시스템 개발 활용 등 문장 검색이 필요한 다른 분야에서도 응용이 가능할 것이다.

참 고 문 헌

[1] 강호정(2007). **과학글쓰기를 잘하려면 기승전결을 버려라**, 이음출판사

[2] 이인철(2007). **동아일보 디지털 스토리 : '표절한국' 이젠 바로잡자**, 동아닷컴

[3] **보도자료 : 표절 문제 전담 해결을 위한 표절 위원회 출범**, 문화체육관광부

[4] 류창건(2008). **일반화된 정렬 가중치 모델을 이용한 내용 기반 문서 검색 시스템**, 한국정보학회 학술발표 논문집

[5] 최성원(2005), **주변 문장 유사도를 이용한 문서 재사용 측정 모델** , 고려대학교 석사 학위 논문

[6] 신동호(2000), **LSA를 이용한 내용기반 검색엔진 시스템** , 서울대학교 석사 학위 논문

[7] 김혜숙(2004), **단어/단어쌍 특징과 신경망을 이용한 두문서간 유사도 측정** , 정보과학회 논문지:소프트웨어 및 응용 제 31권 제 12호

[8] 조정현(2009), **웹 검색과 문서 유사도를 활용한 2단계 신문 기사 표절 탐지 시스템** , 정보처리학회 논문지 B 제 16-B권 제 2호

[9] 장정호(2003), **헬름홀츠머신 학습 기반의 의미 커널을 이용한 문서 유사도 측정** , 한국정보학회 봄 학술발표 논문집 Vol.30

[10] 황인수(2009), **인터넷 검색과 형태소 분석을 이용한 표절검사 시스템의 개발에 관한 연구** , information technology application & management, Vol.16 no.1, pp.21-36

[11] Saket Mengle(2009), **Passage Detection Using Text Classification** , Journal of the american society for information

science and technology, Vol.60 no.4
[12] JangWon Seo(2008), **Local Text Reuse Detection** , SIGIR'08
[13] Paul Clough(2002), **Measuring Text Reuse** , Proceedings of the conference : Association for Computational Linguistics. Meeting , V.40 pp 152-159
[14] Narayanan Shivakumar(1995), **SCAM : A Copy Detection Mechanism for digital Documents**, Proceedings of 2nd International Conference in Theory and Practice of Digital Libraries
[15] 조준희(2009), **한국어 문서 표절 검사를 위한 LSA와 N-gram 기반의 유사 문장 판별**, 고려대학교 석사학위 논문

지 혜 성

2009 한신대학교 소프트웨어학과(학사)
2008~현재 고려대학교 컴퓨터교육과 석사과정
관심분야: 자연어처리,정보검색,컴퓨터교육
E-Mail: hyesung84@korea.ac.kr

조 준 희

1994 경희대학교 기계공학과(학사)
2010 고려대학교 컴퓨터학과 (석사)
20??~현재 (주)유라클 대표이사
관심분야: 컴퓨터교육, 자연어처리
E-Mail: jhhoh@uracle.co.kr

임 희 석

1992 고려대학교 컴퓨터학과 (학사)
1994 고려대학교 컴퓨터학과 (석사)
1997 고려대학교 컴퓨터학과 (박사)
2008~현재 고려대학교 컴퓨터교육과 교수
관심분야: 컴퓨터교육, 자연어처리, 인지신경과학
E-Mail: limhseok@korea.ac.kr