KIPS Trans. Softw. and Data Eng.
Vol.8, No.3 pp.109~114 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2019.8.3.109

Implementation of A Plagiarism Detecting System with Sentence and Syntactic Word Similarities

Joosoo Maeng† · Ji Su Park†† · Jin Gon Shon†††

ABSTRACT

The similarity detecting method that is basically used in most plagiarism detecting systems is to use the frequency of shared words based on morphological analysis. However, this method has limitations on detecting accurate degree of similarity, especially when similar words concerning the same topics are used, sentences are partially separately excerpted, or postpositions and endings of words are similar. In order to overcome this problem, we have designed and implemented a plagiarism detecting system that provides more reliable similarity information by measuring sentence similarity and syntactic word similarity in addition to the conventional word similarity. We have carried out a comparison of on our system with a conventional system using only word similarity. The comparative experiment has shown that our system can detect plagiarized document that the conventional system can detect or cannot.

Keywords : Plagiarism, Partial Plagiarism, Similarity Measurement, Word Similarity, Sentence Similarity, Syntactic Word Similarity, Morpheme Analysis

문장 및 어절 유사도를 이용한 표절 탐지 시스템 구현

맹 주 수† · 박 지 수†† · 손 진 곤†††

요 약

기존 표절 탐지 시스템은 형태소 분석을 기반으로 공통 단어의 빈도수를 이용해 문서의 유사도를 측정한다. 그러나 주제가 같아 유사 단어가 많이 쓰인 경우, 문장 단위로 일부만 발췌 표절한 경우, 그리고 조사와 어미의 유사성이 있는 경우는 공통 단어의 빈도수만으로는 정확한 유사도를 측정하는데 한계가 있다. 따라서 본 논문에서는 공통 단어 빈도수 기반의 유사도 측정 외에 문장 유사도와 어절 유사도를 추가적으로 측정해 유사도의 정확성을 높일 수 있는 표절 탐지 시스템을 설계하고 구현하였다. 실험 결과, 문장 유사도를 측정함으로써 문장 단위로 표절이 이루어진 경우를 발견할 수 있었고, 어절 유사도를 추가로 측정함으로써 부분표절이 일어난 경우라도 조사나 어미까지 그대로 사용한 표절의 경우 등을 발견할 수 있었다.

키워드 : 표절, 부분표절, 유사도 측정, 단어 유사도, 문장 유사도, 어절 유사도, 형태소 분석

1. 서 론

표절은 타인의 독창적인 아이디어 또는 창작물을 적절한 출처 표시 없이 활용하여 자신의 창작물인 것처럼 인식하게 하는 행위이다[1]. 표절에 대한 문제의식이 확대되면서 노골적인 복사 수준의 표절은 더 이상 찾아보기 어렵다. 그러나 문장의 위치 바꾸기, 짜깁기, 어간이나 어미의 변형, 조사나 단어 치환 등 한글의 특징을 이용한 지능적인 표절이 점차 증가하고 있다[2].

표절검사를 위해 개발된 기존의 유사도 측정 시스템들은 대부분 형태소 분석을 통한 공통 단어의 빈도수를 이용해 유사도를 측정한다. 그러나 주제가 같아 유사 단어가 많이 사용된 경우, 해당 분야의 전문 용어가 공통적으로 많이 사용된 경우에는 공통 단어의 빈도수만으로는 정확한 유사도를 측정하는데 한계가 있다[3]. 또한 전체 유사도는 낮으나 부분적으로 조사와 어미의 유사성이 높은 문서의 경우에도 공통 단어 빈도수만으로는 정확한 유사도를 측정하는 것이 불가능하다.

조사나 어미는 개인의 문체를 드러내는 중요한 요소이기 때문에 표절이 일어날 경우 표절여부를 확인할 수 있는 단서가 된다. 그러나 공통 단어 빈도수 기반 유사도 측정 시 형태소 분석을 거치면서 조사와 어미가 원형으로 복원되어 빈도수 정보만 사용된다. 따라서 비교하고자 하는 두 문서에 조사와 어미까지 같은 공통 어절이 몇 개 존재하는지를 추가적으로 측정하면, 기존의 방법으로는 측정하지 못한 새로운 유사도 정

† 준 회 원 : 한국방송통신대학교 이러닝학과 석사과정
†† 비 회 원 : 동국대학교 융합소프트웨어교육원 교수
††† 종신회원 : 한국방송통신대학교 컴퓨터과학과 교수
Manuscript Received : October 8, 2018
First Revision : November 20, 2018
Accepted : November 20, 2018
* Corresponding Author : Jin Gon Shon(jgshon@knou.ac.kr)

보를 얻을 수 있게 된다. 또한 한 문장 단위로 발췌한 곳에서 표절이 일어난 경우에는 전체 문서의 유사도가 낮게 측정될 수 있기 때문에, 정확한 표절 판단을 위해서는 단어 빈도수 기반 유사도 측정 외에 공통 문장이 몇 개 존재하는지 확인할 수 있는 문장 유사도 또한 측정이 필요하다. 문서를 문장단위로 분리하여, 대상문서의 문장들과 문장단위로 비교하여 유사율을 구하고, 포인트를 부여한다. 이 부여된 포인트에 의해서 유사문장이 결정되고, 유사문장이 존재하는 경우 앞서 측정한 유사율들과 같이 유사 문장을 표현해 줌으로써 표절을 판별하는데 있어 부분적인 문장표절에 대해서도 인지할 수 있게 한다.

본 연구는 기존의 문제점을 보완하고자 공통 단어 빈도수와 함께 공통 문장 빈도수 측정을 통해 문장 단위의 부분표절 사례에 대한 추가 발견을 할 수 있고, 공통 어절 빈도수를 추가적으로 측정해 조사나 어미를 그대로 사용한 사례 등을 발견함으로써 기존 방법보다 정확한 유사도를 도출할 수 있는 표절 탐지 시스템을 구현한다.

2. 관련 연구

2.1 형태소 분석

기존 유사도 측정 방법은 형태소 분석을 통해 단어 일치도를 파악하는 방법과 구문-의미 분석을 이용한 방법으로 나누어진다[4, 5]. 이 중 가장 널리 사용되는 것은 형태소 분석을 이용한다.

형태소는 뜻을 가진 가장 작은 단위의 말로써, 형태소 분석은 컴퓨터가 자연어를 처리할 수 있도록 만들어주는 가장 기본적이고 필수적인 과정이다[6]. 형태소 분석은 주어진 문장에서 최소 단위인 형태소를 추출하는 것으로, 문장을 형태소 단위로 분리함으로써 변형이 일어난 형태소의 원형을 복원한 후 품사정보를 추출해 내는 과정을 거친다[6, 7].

2.2 단어 빈도 가중치

단어 빈도(Term Frequency)는 문서 내에서 단어가 사용된 횟수이며, 공통 단어의 출현 빈도는 문서 간 유사도를 측정하는데 중요한 역할을 한다.

tf는 한 단어가 문서내에서 나타난 출현 빈도수이며, TF는 한 단어가 문서내에서 나타난 출현 빈도수의 가중치를 의미한다. 그 중 이진 TF는 단어가 출현한 경우를 모두 1로 지정하여 가중치를 주는 방법이고, 단순 TF는 단어 빈도를 나타내는 TF를 보정 계수 적용 없이 그 횟수만큼 더하여 가중치로 사용하는 것을 의미한다[8]. 이외에도 다양한 가중치와 유사도를 구하는 방법이 존재하나, 본 연구에서는 어절의 어미 등을 활용하여 유사도를 구하는데 중점을 두고 있기에, 형태소 분석 및 어절 분석 과정에서 단순 TF를 적용한다.

2.3 색인과 불용어

색인(indexing)은 방대한 양의 정보원으로부터 정보의 특성을 표현하거나 주제를 나타내는 데이터 요소를 추출하는 작업이다[5]. 색인을 통해 탐색 시간을 최소화하고 빠른 속도로 정보를 제공할 수 있다[6].

불용어(stopword)는 색인 단어로는 의미가 없는 관사, 조사, 접속사 등으로, 형태소 분석의 색인 단어 추출 과정에서는 무시되는 요소이다. 그러나 불용어에 가중치를 두어 유사도를 측정하면, 기존의 공통 단어 빈도수 기반 유사도 측정에 비해 성능이 12% 향상된다[9]. 따라서 본 연구에서는 불용어인 조사와 어미가 사용된 문장에서 어절 단위의 유사도를 추가적으로 조사한다.

3. 표절 탐지 시스템의 유사도 분석 기법

표절 탐지 시스템에서는 공통 단어 빈도수에 따른 유사도 분석을 일차적으로 수행한 후, 공통 문장 빈도수와 공통 어절 빈도수를 추가적으로 조사해 유사도의 정확성을 높인다. 일차적으로 이뤄지는 공통 단어 빈도수 분석에서 유사율이 80% 이상으로 측정된 대응 문서는 표절 의심 문서로 분류된다.

본 논문에서 설정한 80%는 임의 설정 값으로, 유사도 판단 정책에 따라 변경 가능하다. 유사율이 80% 미만으로 측정된 대응문서는 공통 문장 빈도수 분석과, 공통 어절 빈도수 분석을 병렬적으로 시행한다. 표절 탐지 시스템의 전체적인 흐름은 Fig. 1과 같다.

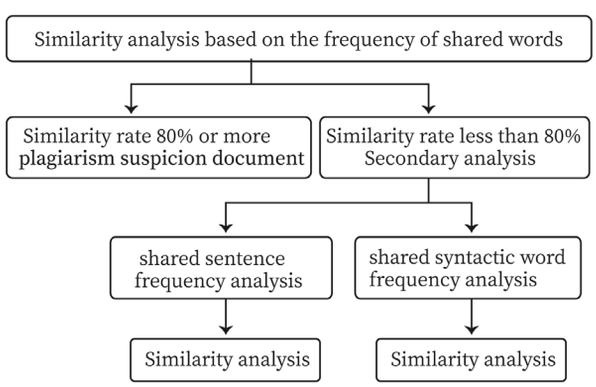


Fig. 1. The Entire System Chart

3.1 공통 단어 빈도수에 따른 유사도 분석

공통 단어 빈도수에 따른 유사도 분석은 입력된 원본 문서와 대응 문서에서 각각 형태소 분석을 통해 색인어를 추출한다. 색인어 추출 시, 각 단어의 품사를 조사하고 단어에 붙은 조사와 어미는 원형으로 복원한다[8]. 원본 문서 기준으로 추출된 색인어들은 대응 문서에서 각각 몇 번의 빈도로 사용되었는지를 단순 TF 방식으로 계산해 단어 빈도수에 따른 유사도를 측정한다. 공통 단어 빈도수에 따른 유사도 분석은 Fig. 2와 같다.

3.2 공통 문장 빈도수에 따른 유사도 분석

공통 문장 빈도수는 원본 문서와 대응 문서의 모든 문장들을 각각 분리하고, 원본 문서에 있는 문장들이 대응 문서에서 그대로 사용된 것이 몇 개인지 확인한다. 공통 문장 빈도수에

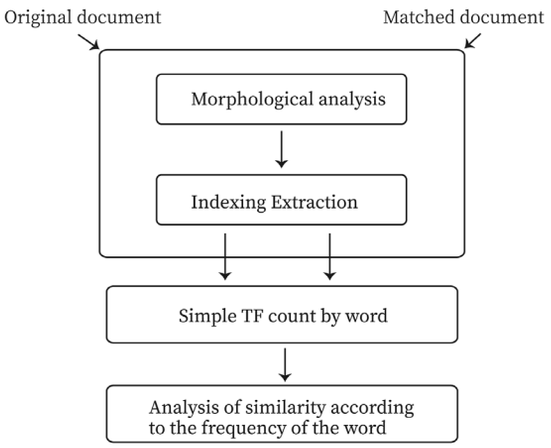


Fig. 2. Similarity Analysis based on the Frequency of Shared Words

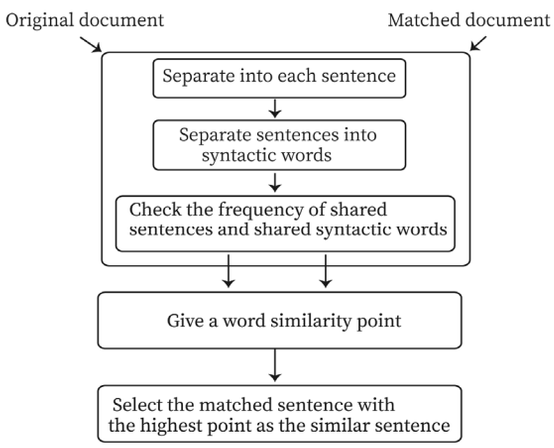


Fig 3. Similarity Analysis based on the Frequency of Shared Syntactic Words

대한 유사율은 공통 문장 수(B)를 대응 문서의 총 문장 수(A)로 나눈 값(B/A)으로 나타낸다.

다음 Table 1에서는 대응 문서 1, 2, 3의 총 문장 수가 각각 10개, 20개, 30개이고, 원본 문서에 있는 문장이 대응 문서에서 각각 5개, 8개, 18개가 사용 되었을 경우의 유사도이다. 유사도에서 3번 대응 문서는 문장 유사율을 측정한 수치가 18/30으로 공통 문장 유사율이 가장 높다.

Table 1. Similarity Analysis based on the Frequency of Shared Sentences

Matched documents number	Total number of sentences in the matched documents (A)	The number of sentences shared with the original document (B)	Similarity (B/A)
1	10	5	5/10
2	20	8	8/20
3	30	18	18/30

3.3 공통 어절 빈도수에 따른 유사도 분석

공통 어절 빈도수에 따른 유사도 분석은 조사와 어미를 그대로 사용한 어절 단위의 유사도 분석을 위해 원본 문서와 대응문서를 각각 문장 단위로 나눈다. 각 문장을 다시 어절 단위로 분리하여 원본 문장의 어절이 대응 문장에 얼마나 나타나는지 빈도수를 측정한다. Fig. 3은 공통 어절 빈도수에 따른 유사도 분석의 절차를 나타낸다.

측정한 빈도수를 이용하여 대응 문장의 유사율을 계산한다. 대응 문장 중 최대의 유사율을 보인 문장을 유사 문장으로 선정하고 이 선정된 유사 문장에 원본 문장의 어절수와 동일한 포인트를 부여한다. 이 포인트를 이용하여 유사 문장을 선정한다. 유사한 두 문서에서의 유사율은 원본 문장과 대응 문장의 공통 어절 빈도수를 각 대응 문장의 총 어절 수로 나누어 계산한다. 공통 어절 빈도수가 동일할지라도 대응 문장의 총 어절 수에 차이가 있다면 유사율은 달라진다. 이중 유사율이 가장 높은 대응 문장에 공통 어절 빈도수만큼 최종 포인트를 준다.

Table 2에서 원본 문장의 어절 수가 3이고 대응 문장 1의 어절수가 3인 경우, 모든 어절이 동일하므로 히트수는 3이며 유사율은 3/3이다. 대응 문장 2는 '병원에'를 뺀 나머지 어절이 모두 동일하기 때문에 히트수는 2가 되고 유사율은 2/3이다. 대응 문장 3은 7개 어절 중에서 '나는', '학교에', '갔다' 등 3개 어절이 동일하므로 히트수는 3, 유사율은 3/7이다. 또한 대응 문장 3은 공통 어절 히트수가 대응 문장 1과 같더라도 원본 문장과 다른 어절을 가지고 있기 때문에 유사율이 대응 문장 1에 비해 작게 나온다. 따라서 가장 비슷한 대응 문장 1에 히트수인 3만큼 최종 포인트를 준다. 최종 포인트를 획득한 대응 문장 1이 원본 문장의 유사 문장으로 선정된다.

Table 2. Syntactic Word Similarity and Final Point

	Sentence	The number of syntactic words	Hit	Similarity	Final point
Original sentence	나는 학교에 갔다.	3	-	-	-
Matched sentence 1	나는 학교에 갔다.	3	3	3/3	3
Matched sentence 2	나는 병원에 갔다.	3	2	2/3	×
Matched sentence 3	나는 학교에 가는 길에 떡볶이를 먹고 갔다.	7	3	3/7	×

4. 표절 탐지 시스템 구현

4.1 개발 환경

시스템 구축을 위한 웹서버는 Apache Tomcat, 유사 문장의 색인단어 저장을 위한 데이터베이스는 MySQL을 사용하며, 사용자 인터페이스 구축을 위해 Servlet/JSP, HTML5, JavaScript, jQuery, CSS3, BootStrap, Chart.js를 이용한다. 본 시스템에서 사용되는 소프트웨어와 하드웨어 환경은 다음 Table 3과 같다.

Table 3. System Environment

Classification			Specifications
Software	Server	OS	Windows 10
		Web Server	Apache Tomcat 8.0.35
	DBMS		MySQL Version 5.7.12
	Language		Servlet/JSP HTML5, JavaScript, jQuery CSS3, Bootstrap, Chart.js
	Language Editor		Eclipse Mars.2
Hardware	CPU		Intel Core i5
	RAM		8GByte
	SSD		512Gbyte

4.2 데이터베이스 설계

표절 탐지 시스템의 전체 흐름은 형태소 분석 후 문장과 어절 분석을 병렬적으로 실행하는 것으로 다음과 같은 순서로 검사한다. 첫 번째는 루씬을 이용하여 각 파일에서 글 데이터를 추출한다. 두 번째는 추출한 원문 데이터에서 형태소를 추출하여 색인 작업을 하고, 세 번째로 대응 문서와 비교하여 유사도를 측정하여 판별한다.

표절 탐지 시스템에서는 다양한 형식의 파일에서 원문 데이터를 미리 추출하여 저장한다. Fig. 4에서 문서 테이블은 원문 데이터를 추출하여 저장하는 테이블이다. 저장된 원문 데이터에서 형태소를 추출하여 형태소 분석 결과 테이블에 저장한 뒤, 형태소의 유사도를 분석한 결과를 형태소 유사도 결과 테이블에 저장한다. 다음으로 문서 테이블에 저장된 본문 데이터를 문장 단위로 분리하여, 문장 테이블에 저장하고, 저장된 문장들을 비교하여 얻어낸 문장 유사도 결과는 문장 유사도 결과 테이블에 저장한다.

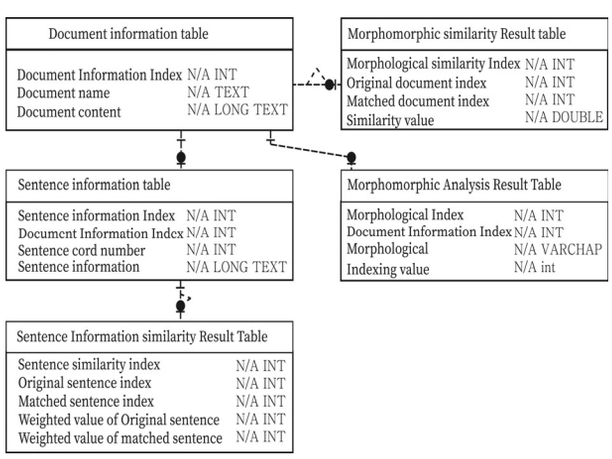


Fig. 4. Database Structure

4.3 표절 의심 문장 탐지 테스트

본 논문에서는 원본 문서 일(一) 대(對) 대응문서 다(多)의 구조로 문서 간의 공통 단어 빈도수를 이용해 유사도를 측정한다. 그 결과, 유사도가 80%이상인 문서는 표절 의심 문서로 분류되고, 유사도가 80% 미만인 대응문서들은 다시 문장 및 어절 단위의 비교를 통해 추가적으로 표절 의심 문서를 구분해 내는 과정을 거친다.

유사 문장 표기 화면은 Fig. 5와 같이 원본 문서와 대응문서를 양쪽에 배치하고, 원본 문서의 문장 중 대응문서와 유사성을 가진 문장에 마우스를 올리면 노란색 하이라이트로 표시한다.

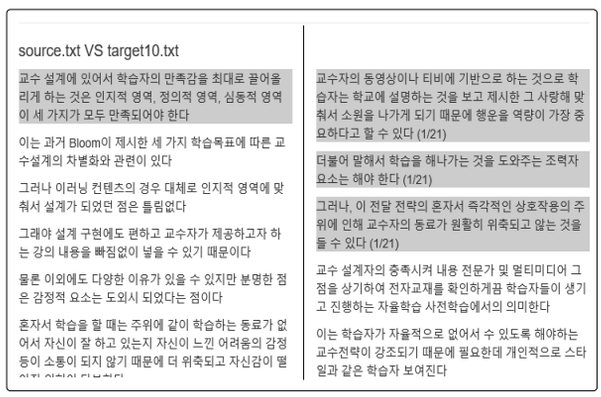


Fig. 5. The Examples of Detecting Similar Sentences

또한 문장이 완전히 동일한 경우에는 Fig. 6과 같이 파란색 하이라이트로 표시하여, 문장단위의 표절이 있을 수 있음을 시각적으로 표시한다.

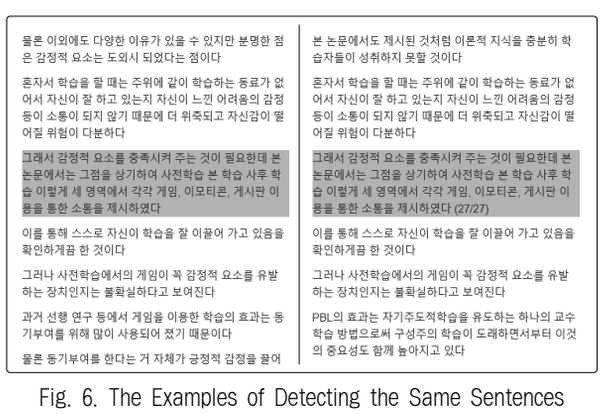


Fig. 6. The Examples of Detecting the Same Sentences

5. 실험 평가

5.1 실험 조건

비교 대상인 M 솔루션은 대용량 통합 검색 솔루션으로 자연어 처리를 이용한 유사도 탐색 시스템이다. M 솔루션은 색인 속도가 1GB당 10분 이내이며, 확장성 및 유연성을 갖춘 JAVA 기반의 검색엔진으로 다양한 DB 색인 및 파일 포맷에 대한 필터가 제공된다[11]. 이 시스템은 형태소를 기반으로 단어 유사도를 이용하며, 본 논문의 공통 단어 빈도수에 따른 유사도 분석 방식과 차이점이 없다. 그러나 제안 시스템에서는 공통 문장 유사도와 공통 어절 유사도를 추가적으로 측정

하고 유사도 측정의 신뢰도를 높이고, 표절 판별을 위해 정확한 정보를 제공한다.

M 솔루션과 제안 시스템 간 유사도 측정의 결과를 비교하기 위해 원본 문서와 10개의 대응문서 사례를 만들어 각각 적용한다. 사례는 Table 4와 같다. 대략 200개정도의 단어로 구성된 한 문서는 비교 기준이 되는 본 문서와 형태소가 유사한 문서, 문장 단위로 유사한 문서, 어절단위로 유사한 문서를 임의로 작성하여 실험한다.

Table 4. Similarities of Matched Documents

Matched documents number	The degree of similarity with an original document
1	Shared 90% of morpheme with an original document
2	Shared 80% of morpheme with an original document
3	Shared 8 sentences with an original document
4	Shared 6 sentences with an original document
5	Shared 5 sentences with an original document
6	Shared 3 sentences with an original document
7	Changed the order of syntactic words
8	Different from an original document, but shared 46 syntactic words
9	Different from an original document, but shared 40 syntactic words
10	Different from an original document, but shared 30 syntactic words

5.2 비교 평가

원본 문서와 10개의 대응문서를 두 시스템에서 분석한 결과는 Table 5와 같다.

Table 5. Word Similarity Result Comparison

Matched documents number	Similarity results of M Solution[%]	Similarity results of Plagiarism Detecting System[%]	Plagiarism Suspicion Document
1	95	84	O
2	94	81	O
3	82	83	O
4	67	73	-
5	74	76	-
6	36	45	-
7	99	99	O
8	44	58	-
9	41	56	-
10	36	49	-

M 솔루션의 경우는 형태소 분석에 따른 단어 유사도만 제시되며, 대응문서 7이 유사도가 가장 높고(99%) 대응문서 10이 유사도가 가장 낮다(36%). 제안 시스템의 결과는 대응문서 7이 유사도가 가장 높고(99%) 대응문서 6이 유사도가 가장 낮은(45%) 것으로 측정됐다. 대응문서 2의 유사도(M 솔루션 94%, 표절 탐지 시스템 81%)를 제외하면 큰 차이는 없다.

대응문서 1, 2, 3, 7은 양쪽의 단어 유사도 결과에서 모두 유사율이 80% 이상으로 측정되어 표절 의심 문서로 분류된다. 따라서 이를 제외한 대응문서 4, 5, 8, 9, 10을 대상으로 문장 유사도와 어절 유사도를 추가적으로 측정한다.

문장유사도 측정 결과는 Table 6과 같다. 대응문서 4의 경우 전체 11문장 중 7문장이, 대응문서 5의 경우 11문장 중 8문장이 원본 문장과 동일해 추가적으로 표절 의심 문서로 분류한다. 특히 대응문서 6의 경우, 단어 유사도는 가장 낮게 나왔으나, 원본 문서와 동일한 3문장이 발견되어 표절에 대한 검사가 따로 필요하다. 8, 9, 10번 대응문서에서는 동일한 문장이 발견되지 않았다.

Table 6. Sentence Similarity

Matched documents number	Sentence similarity
4	7/11
5	8/11
6	3/10
8	0/10
9	0/10
10	0/10

어절 유사도 결과는 Table 7과 이 대응문서 5가 가장 높고(73%) 대응문서 8과 대응문서 10이 가장 낮다(20%). 대응문서 5, 대응문서 4의 경우 원본 문서와 동일한 문장을 7~8개 포함하고 있기 때문에 어절 유사도에서도 순위가 높게 측정된다.

Table 7. Syntactic Word Similarity

Matched documents number	Syntactic word similarity[%]
4	68
5	73
6	33
8	20
9	22
10	20

이와 같이 세 가지 유사도 검사를 실시한 결과, 단어 빈도수에 따른 유사도를 통해 표절 의심 문서로 분류된 4개의 문서 외에도 추가적으로 표절 의심 문서로 분류할 수 있는 문서를 3개 더 탐지해 낼 수 있다는 것을 확인하였다.

6. 결 론

일반적으로 이루어지는 단어 기반 유사도 측정 방법은 형태소 분석을 기반으로 공통 단어의 빈도수를 이용해 문서의 유사도를 측정한다.

본 논문에서는 일차적으로 단어 빈도수 기반 유사도를 측정해, 기준보다 높은 유사도를 보이는 문서에 대해서는 표절 의심 문서로 분류한다. 그 다음 기준보다 낮은 유사도를 보이

는 문서에 대해서는 다시 문장 유사도 분석과 어절 유사도 분석을 추가적으로 실시하여 보다 신뢰성 있는 유사도를 도출할 수 있도록 한다. 그 결과 기존시스템으로는 찾아내지 못했던 표절 의심 문서를 추가적으로 발견한다.

이렇게 단어 기반, 문장 기반, 어절 기반의 3가지 유사도 분석을 통해 보다 신뢰할 수 있는 유사도 검사가 이루어지면 부분 표절 등 다양해져가는 표절에 상응하는 표절 탐지가 이루어진다.

향후 연구 과제로 유사도 측정의 정확성을 향상시키기 위한 구문분석, 의미 분석이 가능한 시스템의 개발이 필요하다. 또 유사어, 동의어, 상의어, 하의어 등 단어 간 포함관계를 반영한 사전 구축 및 복문을 인식할 수 있는 알고리즘 개발이 요구된다[4, 8, 12].

References

[1] Ministry of “Education, Instructions to Securing Research Ethics,” 2015.

[2] Jun, M. J, Park, S. D., Park W., Heo, J. Y., and Cho, H. G., “Plagiarised Reports Detection System using Characteristcs of Korean Language and Local alignment Algorithm,” *Journal of KIISE,* Vol.31, No.02, pp.727-729, 2004.

[3] Seung-hee Yoo, Yil-hyeong Mun, and Dong-sub Cho, “Similarity Measurement of Korean Documents using the Specified Particles and Major Keywords,” *Journal of Korea Multimedia Society*, Vol.2007, No.1, pp.0686-0688, 2007.

[4] Sang Wook Park, Jeong Yoon Kim, Tae Hoon Lee, Seung Beom Hong, Jin Sook Lim, and Won Seog Kang, “Development of Document Plagiarism Detection Algorithm using Syntactic Analysis Method,” *The Korean Association of Computer Education*, Vol.17, No.1, pp.89-93, 2013.

[5] Bang-Won Ko and Young-Chul Kim, “A Similarity Valuating System using The Pattern Matching,” *Journal of the Korea Society of Computer and Information*, Vol.15, No.1, pp.185-192, 2010.

[6] J. H. Choi and S. J. Lee, “A Method for Reducing Dictionary Access with Bidirectional Longest Match Strategy in Korean Morphological Analyzer,” *Journal of KIISE*, Vol.20, No.10, pp.1497-1507, 1993.

[7] Kang Seung-Shik, “Multi-level Morphological Analysis Model for Korean,” *Journal of KIISE*, Vol.1994, No.10, pp.140-145, 1994.

[8] Lee Mi-suk, “A copy detection system,” Ph.D. dissertation, University of Dongguk, Seoul, Korea, 2005.

[9] Won Ji Hur and Yong Gyu Jung, “A Study on Improved Measurement of Similarity Between Documents,” *Journal of KIISE*, Vol.38, No.2, pp.122-124, 2011.

[10] Erik Hatcher, Otis Gospodnetic, and Mike McCandless, “Lucene in Action,” pp.68-69, 2010.

[11] Diquest Mariner2 [internet], http://cfile248.uf.daum.net/image/2509DF40552DACBE05C48A. 2018. 11. 18

[12] Go Eun-byeol, “String and Sentence Similarity Measurement Methods Using Set-based POI Search Algorithm,” Ph.D. dissertation, Sookmyung Women's University, Seoul, Korea, 2014.

맹 주 수

https://orcid.org/0000-0002-0820-2964

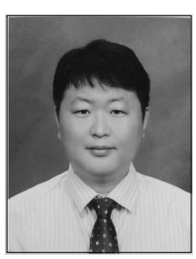
e-mail : ddray@knou.ac.kr

2008년 숭실대학교 소프트웨어공학과 (공학석사)

2018년～현 재 한국방송통신대학교 이러닝학과 석사과정

2013년～현 재 ㈜월비소프트 대표이사

관심분야 : 유사도, 표절, 검색엔진, e-Learning



박 지 수

https://orcid.org/0000-0001-9003-1131

e-mail : bluejisu@dgu.edu@knou.ac.kr

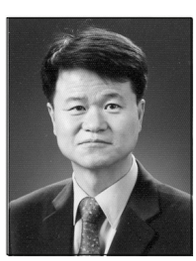
2013년 고려대학교 컴퓨터교육과 (이학박사)

2015년～2018년 충남대학교 초빙교수

2019년 경기대학교 교양학부 조교수

2019년～현 재 동국대학교 융합소프트웨어교육원 교수

관심분야 : 분산 시스템, 모바일 클라우드 컴퓨팅, e-Learning



손 진 곤

https://orcid.org/0000-0002-0540-4640

e-mail : jgshon@knou.ac.kr

1991년 고려대학교 전산학전공(이학박사)

1991년～현 재 한국방송통신대학교 컴퓨터과학과 교수

1997년～1998년 State University of New York (Stony Brook) Visiting Professor

2000년～현 재 ISO/IEC JTC1/SC36 Korea Delegate

2010년 한국정보처리학회 부회장

2009년～현 재 이러닝학회 부회장

관심분야 : 컴퓨터통신망, 분산시스템, e-Learning, 정보기술 표준화