Journal of the Korea Industrial Information Systems Research Vol. 22 No.2, Apr. 2017 :43-52
http://dx.doi.org./10.9723/jksiis.2017.22.2.043 ISSN:1229-3741

이산 푸리에 변환을 적용한 텍스트 패턴 분석에 관한 연구 - 표절 문장 탐색 중심으로 -

(A Study on Text Pattern Analysis Applying Discrete Fourier Transform - Focusing on Sentence Plagiarism Detection -)

이 정 송[1], 박 순 철[2]*
(Lee Jung-Song and Park Soon-Cheol)

요 약 패턴 분석은 신호 및 영상 처리와 텍스트 마이닝 분야에서 가장 중요한 기술 중 하나이다. 이산 푸리에 변환(Discrete Fourier Transform: DFT)은 일반적으로 신호와 영상의 패턴을 분석하는데 사용된다. 본 논문에서는 DFT가 텍스트 패턴 분석에도 적용될 수 있음을 가정하고 문서의 텍스트 패턴이 다른 문서에서도 존재하는지를 탐색하는 표절 문장 탐색에 세계 최초로 적용하였다. 이를 위해 텍스트를 ASCII 코드로 변환하여 신호화하고 복사/붙여넣기, 용어의 재배치 등 단순한 표절 형태의 탐색은 Cross-Correlation(상호 상관)을 이용하였다. 또한 유의어를 사용하거나 번역 및 요약 등의 표절 형태를 탐색하기 위해 워드넷(WordNet) 유사도를 사용하였다. 실험을 위해 표절 탐색 분야의 저명한 워크숍인 PAN에서 제공하는 공식적인 데이터 셋(2013 Corpus)을 사용하였으며, 실험 결과 11개의 표절 문장 탐색 기법 중 4번째로 우수한 성능을 보였다.

핵심주제어 : 이산 푸리에 변환, 표절 문장 탐색, 텍스트 신호화, 상호 상관, 워드넷

Abstract Pattern Analysis is One of the Most Important Techniques in the Signal and Image Processing and Text Mining Fields. Discrete Fourier Transform (DFT) is Generally Used to Analyzing the Pattern of Signals and Images. We thought DFT could also be used on the Analysis of Text Patterns. In this Paper, DFT is Firstly Adapted in the World to the Sentence Plagiarism Detection Which Detects if Text Patterns of a Document Exist in Other Documents. We Signalize the Texts Converting Texts to ASCII Codes and Apply the Cross-Correlation Method to Detect the Simple Text Plagiarisms such as Cut-and-paste, term Relocations and etc. WordNet is using to find Similarities to Detect the Plagiarism that uses Synonyms, Translations, Summarizations and etc. The Data set, 2013 Corpus, Provided by PAN Which is the One of Well-known Workshops for Text Plagiarism is used in our Experiments. Our Method are Fourth Ranked Among the Eleven most Outstanding Plagiarism Detection Methods.

Key Words : Discrete Fourier Transform, Sentence Plagiarism Detection, Text Signal, Cross-Correlation, WordNet

* Corresponding Author : scpark@jbnu.ac.kr
Manuscript receivrd Feb, 3, 2017 / revised Mar, 3, 2017 / accepted Mar, 9, 2017
1) 전북대학교 전자정보공학부, 제1저자
2) 전북대학교 컴퓨터공학부, 교신저자

1. 서 론

신호 처리 분야에서 가장 중요한 이론 중 하나인 푸리에 변환(Fourier Transform)은 시간 영역(Time Domain)의 신호를 주파수 영역(Frequency Domain)으로 변환하여 신호의 주파수 성분 분석에 필수적으로 사용된다. 디지털 신호를 분석할 때는 이산 푸리에 변환(Discrete Fourier Transform)이 사용되어지며 계산속도 향상을 위해 고속 푸리에 변환(Fast Fourier Transform)이 개발되었다[1]. 또한, 영상 처리에서 공간 영역(Spatial Domain)의 주파수 영역 변환을 통해 활용되고 있으며 이 뿐만 아니라 다양한 연구 분야에도 적용되고 있다[2-4].

신호 및 영상 처리에서 유사한 음성 신호를 찾거나 이미지에서 특정 객체를 찾는 템플릿 매칭(Template Matching)처럼 패턴 분석(Pattern Recognition)을 위하여 이산 푸리에 변환이 기본적으로 사용되고 있다[5]. 패턴 분석은 신호 및 영상 처리뿐만 아니라 텍스트 데이터 마이닝에서도 중요하다. 텍스트 데이터 마이닝에서 패턴 분석은 문장, 문단 또는 문서의 유사성을 측정하여 문서 군집화(Document Clustering), 요약(Document Summarization) 또는 표절 탐색(Plagiarism Detection) 등 다양한 기법의 필수 요소이다. 하지만 이산 푸리에 변환이 패턴 분석에 적용된다는 사실을 기반으로 텍스트 데이터 마이닝에도 적용될 수 있는 가능성이 있음에도 불구하고 현재까지 적용된 연구 사례가 없다.

본 논문에서는 세계 최초로 이산 푸리에 변환을 적용한 텍스트 패턴 분석 기법을 제안한다. 텍스트를 신호화하는 두 가지 방법(ASCII 코드, 워드넷 유사도)을 기반으로 다양한 실험을 진행하였으며 그 결과, 이산 푸리에 변환은 텍스트 문서에서 유사한 패턴을 탐색할 수 있다는 점을 발견하였다. 또한, 다양한 연구 분야 중 표절 문장 탐색에 적용하여 제안하는 기법의 실험적 입증을 보이고자 한다.

인터넷의 등장과 네트워크 발달은 각 사용자들의 시공간적 제약을 줄어들게 만들었고 전 세계 모든 사람들을 연결되게 만들었다. 이것은 각 사용자들에 의해 생성, 공유되어지는 데이터를 기하급수적으로 늘어나게 만드는 요인이 되었다. 이러한 변화 속에서 빅데이터 분석 등 사회에 좋은 효과를 주는 반면에 악효과를 주는 면도 다양하게 존재하는데 그중에서 대표적인 예가 텍스트 표절이다 [6-7]. 다양한 형태의 텍스트 표절을 사람이 직접 판단하기에는 매우 난해하다. 현재까지 텍스트 표절 탐색을 위해 다양한 연구가 진행되고 있으며 가장 일반적으로 사용하는 기법은 지문법(fingerprint)[8]이다. 지문법은 텍스트 내의 사용된 단어의 개수, 단어의 빈도수 등 특징을 추출하여 유사성을 측정하고 표절 유무를 판단한다. 하지만 이러한 기법의 가장 큰 단점은 유의어를 사용하거나 번역 및 요약 형태의 표절 탐색 성능이 현저히 떨어진다는 것이다. 이를 위해 본 논문에서는 이산 푸리에 변환을 기반으로 하여 Cross-Correlation(상호 상관)을 이용한 표절 문장 탐색 기법과 유의어, 변역 및 요약 형태의 표절 탐색을 위해 워드넷 유사도를 사용한 새로운 텍스트 신호화를 이용한 표절 문장 탐색 기법을 제안한다.

본 논문의 구성은 2장에서는 이산 푸리에 변환에 대한 설명과 텍스트 신호화에 대하여 논의하고 3장에서는 이산 푸리에 변환을 적용한 표절 문장 탐색 기법(Cross-Correlation을 이용한 표절 문장 탐색, 워드넷 유사도 기반 새로운 텍스트 신호화를 통한 표절 문장 탐색)을 제안한다. 마지막으로 4, 5장에서는 실험 결과 분석, 결론 및 향후 연구에 대하여 서술한다.

2. 이산 푸리에 변환과 텍스트 신호화

푸리에 변환의 기본 개념은 모든 신호는 각기 다른 주파수를 가진 정현파들의 합으로 표현할 수 있으며 아날로그 연속 신호에 대한 푸리에 변환식과 역변환식은 다음과 같다.

$$X(\omega) = \int_{-\infty}^{+\infty} x(t) e^{-j\omega t} dt \quad (1)$$

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} X(\omega) e^{j\omega t} d\omega \quad (2)$$

자연 현상에 존재하고 측정되는 값들은 모두 연속적인 값들로 표현되는 아날로그 형태이다. 이를 컴퓨터로 계산하기 위해서는 일정한 간격의 값으로 표현될 수 있는 디지털화가 필요하다. 변환된 디지털 신호의 분석을 위해 이산 푸리에 변환이 사용되며 다음 식과 같다.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi kn}{N}} \quad (3)$$

신호 및 영상 처리에서 이산 푸리에 변환은 Cross-Correlation(상호 상관), Auto-Correlation (자기 상관) 등과 함께 이용하여 음성의 유사한 패턴 탐색, 이미지에서 특정 부분 탐색(템플릿 매칭:Template Matching) 등 패턴 분석 분야에 다양하게 활용되고 있다[9]. 따라서 본 논문에서는 패턴 분석에 활용되고 있는 이산 푸리에 변환을 텍스트 패턴 분석에 적용하고자 한다.

이를 위해서 우선적으로 자연어로 쓰인 텍스트 문서의 용어들을 수치화하는 텍스트 신호화가 필요하다. 본 장에서는 일차적으로 텍스트 문서에서 특수문자(마침표, 물음표, 따옴표 등)를 제외한 모든 문자를 ASCII 코드 값으로 변환하는 텍스트 신호화 방법을 제안하며 다음과 같다.

S = {C_1, C_2, ..., C_i, ..., C_n}와 같이 n개의 문자 C_i로 이루어진 문장 S가 있다고 가정하면 S는 각 문자 C_i의 ASCII 코드 값(CA_i)으로 구성된 S' = {CA_1, CA_2, ..., CA_i, ..., CA_n}로 변환된다. 최종적으로 S'는 정규화된 D = {D_1, D_2, ..., D_i, ..., D_n}로 신호화 된다. 여기서 D_i는 식 (4)와 같다.

$$D_i = \frac{CA_i - E(S')}{\sigma_{S'}} \quad (4)$$

여기서, $E(S')$와 $\sigma_{S'}$는 아래와 같다.

$$E(S') = \frac{1}{n}\sum_{i=1}^{n} CA'_i \quad (5)$$

$$\sigma_{S'} = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(CA'_i - E(S'))^2\right)} \quad (6)$$

예를 들면, 아래의 2개의 문장을 가지고 있는 S는 S'와 같이 ASCII 코드 값으로 변환되며 식 (4)를 통하여 최종적으로 신호화된 D로 구성된다.

S = {The Lectro-Kennel dog house heaters are a simple, affordable way to keep your pet comfortable. Constructed of rugged ABS plastic with a steel wrapped cord, the Lectro-Kennel can lie flat on the dog house floor or can be attached to the wall.}
S' = {116, 104, 101, ..., 97, 108, 108}
D = {8.36, -3.63, -6.63, ..., -10.63, 0.36, 0.36}

3. 이산 푸리에 변환을 적용한 텍스트 문서에서의 표절 문장 탐색

앞 절에서 논의한 텍스트 신호화를 기반으로 다양한 실험을 한 결과, 이산 푸리에 변환이 텍스트 데이터 마이닝에 적용 되었을 경우 두 가지 특징을 가지는 것을 발견할 수 있었으며 다음과 같다.

① **유사한 패턴을 가지는 텍스트 탐색**: 동일한 용어들로 이루어진 같은 길이의 텍스트들과 그렇지 않은 텍스트들의 집합 속에서 텍스트들을 신호화하고 이산 푸리에 변환을 적용하였을 경우, 유사한 패턴을 가진 텍스트들의 유사도가 높게 측정되었다.

② **동일한 위치에 동일한 용어를 가지는 텍스트 탐색**: 유사한 패턴을 가지면서도 동일한 위치에 동일한 용어를 가지는 텍스트는 다른 텍스트보다 더욱 높은 유사도를 가짐을 알 수 있었다.

본 논문에서는 이러한 실험 결과를 토대로 이산 푸리에 변환이 텍스트 데이터 마이닝에 적용되어 질 수 있음을 확인하였고 표절 탐색 분야에 응용하여 더욱 확실한 입증을 보이고자 한다. 이산 푸리에 변환을 적용한 표절 문장 탐색 절차는 크게 Cross-Correlation(상호 상관)을 이용한 표절 문장 탐색과 워드넷 유사도 기반 신호화에 푸

리에 변환을 이용한 표절 문장 탐색으로 구성되어진다.

3.1 Cross-Correlation(상호 상관)을 이용한 표절 문장 탐색

본 논문에서 제안하는 신호화 방법을 통해 텍스트를 신호화 하였을 경우, 신호의 유사성을 판별하기 위해 신호 처리 분야에서 두 개의 신호의 유사도를 측정하는데 가장 기본적으로 사용하는 Cross-Correlation(상호 상관) 기법을 적용하였다. 이를 이용하여 표절 문장을 탐색하는 알고리즘은 다음과 같으며 Fig. 1은 실제 아래의 알고리즘을 통하여 표절 문장을 탐색 후 표절로 판단된 예를 나타낸다.

Step 1: 표절 의심 문장과 원본 문장 ASCII 코드 기반 신호화

Step 2: 표절 의심 문장 신호(**signalSusp**)와 원본 문장 신호(**signalSrc**) Cross-Correlation

Step 3: 각각의 신호 재구성

Step 3-1: maxCoffIndex ← 두 신호의 Cross-Correlation 값에서 가장 큰 값의 인덱스
lagsArray ← [-(신호최대길이-1) to (신호최대길이-1)]

shiftIndex ← lagsArray[maxCoffIndex]

Step 3-2: 각각의 신호를 shiftIndex의 절댓값부터 각 길이의 끝까지로 재구성

Step 3-3: 두 신호의 길이를 맞추기 위한 각각의 신호 제로 패딩

Step 4: 재구성된 두 신호(**signalSusp'**, **signalSrc'**) 고속 푸리에 변환

Step 5: 고속 푸리에 변환 후 각 신호(**signalSuspFFT**, **signalSrcFFT**)의 Magnitude 값으로 코사인 유사도 계산

Step 6: **if** 코사인 유사도 값 > 기준값

표절 문장으로 판단

else

워드넷 유사도 기반 신호화와 푸리에 변환을 이용한 표절 문장 탐색

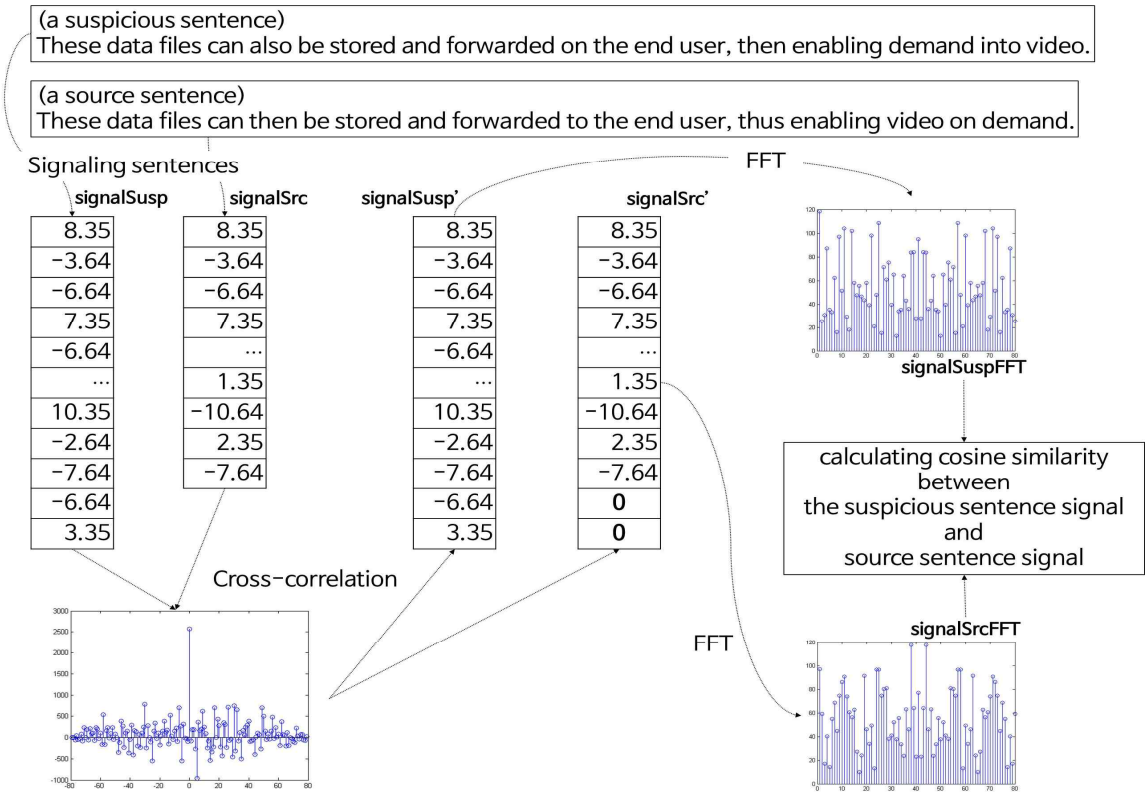


Fig. 1 Sentence Plagiarism Detection using Cross-Correlation

요약하면, 신호화된 표절 의심 문장(**signalSusp**)과 원본 문장(**signalSrc**)을 Cross-Correlation을 적용 한 후 가장 큰 Cross-Correlation 계수 값의 인덱스를 기준으로 신호를 잘라내어 재구성한다. 위의 예에서는 가장 큰 Cross-Correlation 계수 값이 0이기 때문에 신호를 잘라냄 없이 문장의 길이를 맞추기 위해 제로 패딩[10]만 하였음을 볼 수 있다. 다음으로 재구성된 각 신호(**signalSusp'**, **signalSrc'**)에 고속 푸리에 변환을 적용한 후 두 신호(**signalSuspFFT**, **signalSrcFFT**)의 각 주파수 대역의 크기(Magnitude)를 특징으로 하여 코사인 유사도를 측정한다. 마지막으로 코사인 유사도가 기준값 이상일 때 표절 문장으로 판단하고 그렇지 않은 경우에는 다음절에 논의할 새로운 신호화를 기반으로 하는 이산 푸리에 변환을 이용한 표절 문장 탐색을 수행한다.

3.2 워드넷 유사도 기반 텍스트 신호화와 이산 푸리에 변환을 이용한 표절 문장 탐색

워드넷이란 1985년 Princeton 대학의 Miller에 의해 용어들의 개념과 의미적인 유사도, 연관성을 분석하기 위해 개발 된 것으로 텍스트 데이터 마이닝 등 다양한 분야에 활용되어 지고 있다[11]. 워드넷은 명사, 동사, 형용사, 부사 각 품사의 어휘들을 synset이라는 동의어 집합으로 구조화 되어 있으며 이들은 IS-A 계층적 관계를 이룬다. synset은 hypernym(상위어), hyponym(하위어) 등 의미 관계를 제공하며 기본적으로 워드넷을 이용한 두 용어의 의미적 유사도는 상하위어를 지나는 최소 거리로 정의되며 이 외에도 다양한 방법들이 제안되고 있다. 본 논문에서는 WUP[12], RES[13], JCN[14], LCH[15], LIN[16], LESK[17] 총 6개의 다양한 워드넷 유사도 기법을 사용하였다.

Cross-Correlation을 이용한 표절 문장 탐색은 단순히 일치하거나 단어의 순서가 바뀐 문장만을 탐색하기 때문에 높은 성능을 기대하기 어렵다. 그 이유는, Cross-Correlation을 이용한 표절 문장 탐색에서의 문장 신호화는 단순히 각 문자를 ASCII 코드 값으로 변환하기 때문에 의미상으로 유사한 용어를 사용한 표절 문장을 탐색하지 못하기 때문이다. 이를 위해서 이번 절에서는 워드넷 유사도를 기반으로 하는 새로운 문장 신호화 기법을 제안하며 아래와 같이 정의 할 수 있다.

$$signal_{susp} = WN\{T_{susp} \times T_{union}\} \quad (7)$$

$$signal_{src} = WN\{T_{src} \times T_{union}\} \quad (8)$$

여기서, $signal_{susp}$, $signal_{src}$는 각각 T_{susp}, T_{src}와 T_{union}의 곱집합 형태로 신호화된 표절 의심 문장, 원본 문장을 의미한다. 그리고 T_{susp}, T_{src}는 각각 표절 의심 문장, 원본 문장에 포함된 용어 집합을 나타내며 T_{union}는 T_{susp}, T_{src}집합의 합집합을 의미한다. 즉, T_{union}는 표절 의심 문장과 원본 문장에 포함된 전체 용어 집합을 의미한다. WN은 워드넷 유사도를 나타내며 결론적으로 곱집합 형태의 용어쌍들의 모든 워드넷 유사도를 계산하여 표절 의심 문장과 원본 문장을 신호화한다.

워드넷 유사도 기반 텍스트 신호화와 이산 푸리에 변환을 이용한 표절 문장 탐색 절차는 다음과 같다.

Step 1: T_{susp} ← 표절 의심 문장에 포함된 용어
T_{src} ← 원본 문장에 포함된 용어
T_{union} ← $T_{susp} \cup T_{src}$
similarityAVG ← 0

Step 2: for t **in** T_{union}
용어 t와 표절 의심 문장에 포함된 모든 용어와의 워드넷 유사도를 통한 신호화
용어 t와 원본 문장에 포함된 모든 용어와의 워드넷 유사도를 통한 신호화
두 신호의 길이를 맞추기 위한 각각의 신호 제로 패딩
제로 패딩 된 두 신호 고속 푸리에 변환
similarityAVG ← similarityAVG + 고속 푸리에 변환 후 각 신호의 Magnitude값으로 코사인 유사도 계산

Step 3: if similarityAVG / T_{union}의 용어 개수 > 기준값
표절 문장으로 판단

4. 실험 및 분석

본 논문에서 제안하는 이산 푸리에 변환을 적용한 표절 문장 탐색에 대한 성능 실험을 위해 PAN(http://pan.webis.de)에서 제공하는 데이터 셋을 사용하였다. PAN은 표절 탐색 분야에서 가장 저명한 워크숍으로써 표절 검사, 소스 코드 복제, 작가 확인 및 프로파일링 등의 공모전을 개최한다. 본 논문에서 사용한 데이터 셋 "pan13-text-alignment-test-corpus1-2013-03-08"은 398개의 표절 의심 문서와 489개의 원본 문서로 구성되어 있다. 그리고 no-plagiarism, no-obfuscation, random-obfuscation, translation-obfuscation, summary-obfuscation과 같이 총 5개의 표절 형태로 구성되어진 518개의 표절 탐색 결과 정답셋을 포함한다. no-plagiarism은 표절로 판단되는 문장이 없는 경우, no-obfuscation은 cut-and-paste copying 형태의 단순한 표절, random-obfuscation은 단어 교체 및 재배치, 단어 또는 문단 추가 및 삭제 형태의 표절, translation-obfuscation은 자동 번역기를 이용하여 영어를 다른 언어로 번역한 뒤 이를 다시 영어로 번역한 형태의 표절, 그리고 summary-obfuscation은 문단을 요약한 형태의 표절을 나타낸다[18].

본 논문에서는 다음과 같은 실험 단계를 통해 제안하는 이산 푸리에 변환을 적용한 텍스트 문서에서의 표절 문장 탐색 성능을 입증하고자 한다.

① Cross-Correlation 기법만을 이용하였을 경우 표절 문장을 판단하는 기준값에 따른 성능 비교
② 이산 푸리에 변환을 적용하였을 경우 선택되는 주파수 대역에 따른 성능 비교
③ 워드넷 유사도 기반 새로운 문장 신호화에 이산 푸리에 변환을 적용하였을 경우 성능 비교

또한, 실험을 위해 제안하는 기법은 **JAVA**언어를 사용하여 구현하였으며 세부적으로 문장 분리와 용어 추출은 **Stanford CoreNLP 3.6.0** 그리고 워드넷은 **WordNet 3.0**을 사용하였다. 성능을 측정하기 위해서는 표절 탐색 분야에서 가장 많이 쓰이는 Precision, Recall, 그리고 Granularity, 이 3가지 기법을 혼합한 Plagiarism Detection Score(**plagdet**)를 사용하였다 [19-23].

4.1 기준값에 따른 성능 비교

3.1 절에서 논의한바 같이 Cross-Correlation을 이용하였을 경우 표절 문장의 여부를 판단하기 위해 기준값이 필요하다. 즉, 표절 의심 문장과 원본 문장 간의 코사인 유사도가 기준값 이상일 때 표절 문장으로 결정된다. 여기서, 기본적으로 문장은 다수의 공백을 포함하므로 공백을 제거하지 않고 문장을 신호화하였을 경우 큰 노이즈가 발생한다. 따라서 본 논문에서는 공백을 포함한 문장과 제거한 문장, 2개의 문장을 모두 신호화하여 표절 문장 탐색을 수행하였으며 결론적으로 2개의 기준값이 필요하다.

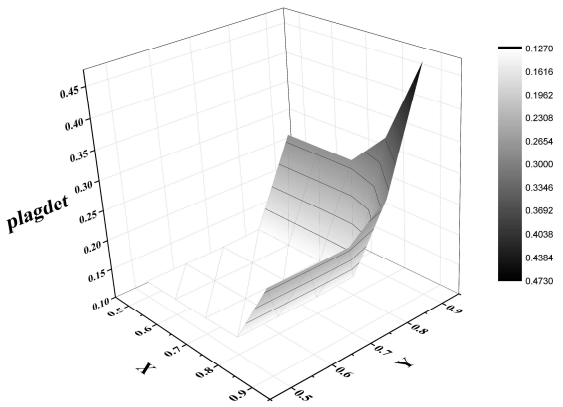


Fig. 2 Changes in Performance according to Parameter (X, Y) in Sentence Plagiarism Detection using Cross-Correlation

Fig. 2는 2개의 기준값(X: 공백 포함, Y: 공백 제거)에 따른 Plagiarism Detection Score (**plagdet**)를 나타낸다. Fig. 2에서 표절 문장을 판단하는 기준값이 0.9, 0.9 (각 공백을 포함한 문장의 기준값, 공백을 제거한 문장의 기준값)일 때 가장 높은 **plagdet**를 가짐을 알 수 있다. 따라서 두, 세 번째 실험의 Cross-Correlation을 이용한 표절 문장 탐색 단계에서는 기준값을 0.9, 0.9로 설정하였다.

4.2 주파수 대역 선택에 따른 성능 비교

신호화된 문장에 이산 푸리에 변환을 적용하였을 경우 주파수 대역은 문장의 길이에 따라 달라진다. 따라서 본 실험에서는 이산 푸리에 변환 후 선택되어지는 주파수 대역에 따른 성능을 비교하였다. 주파수 대역은 0Hz부터 10%~100%의 비율로 10% 간격으로 선택하여 실험하였으며 그 결과는 Fig. 3과 같다.

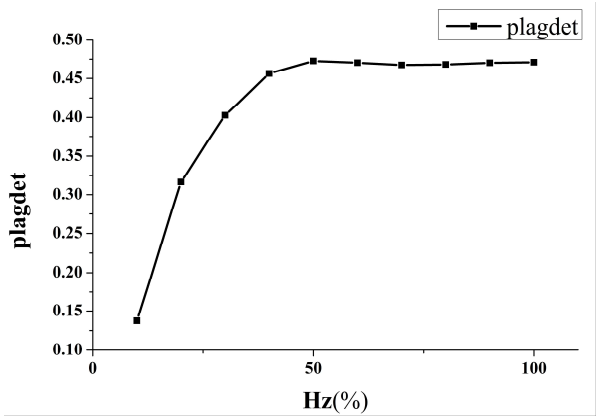


Fig. 3 Changes in Performance in Sentence Plagiarism Detection according to the Selected Frequency Band

Fig. 3에서와 같이 50%의 주파수 대역을 선택하였을 때 가장 높은 성능을 보였으며 그 후 부터는 일정한 성능을 보였다.

4.3 새로운 문장 신호화에 따른 성능 비교

Cross-Correlation을 이용한 표절 문장 탐색에서는 유의어를 사용한 표절 문장 탐색 성능이 낮기 때문에 전체적으로 높은 성능을 기대하기 어려웠다. 따라서 본 논문에서는 워드넷 유사도 기반 새로운 문장 신호화를 제안하고 이를 위해 6개(WUP, RES, JCN, LCH, LIN, LESK)의 워드넷 유사도 기법을 사용하였다. Fig. 4는 각 6개의 워드넷 유사도 기반 문장 신호화를 이용한 표절 문장 탐색 성능을 비교한 것이며 LESK 유사도 기법이 가장 높은 성능을 보였다.

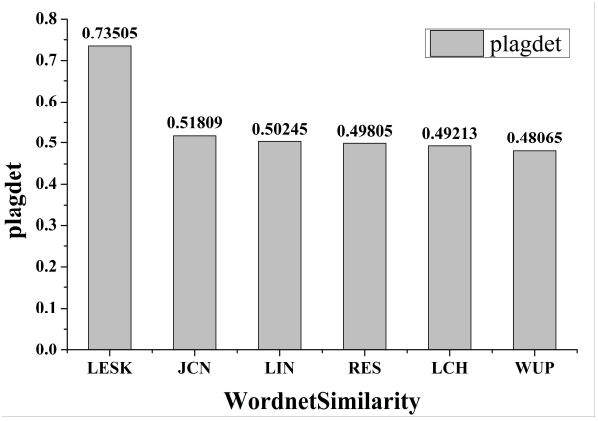


Fig. 4 Changes in Performance in Sentence Plagiarism Detection according to WordNet Similarity Measure

또한, 2013년도에 PAN 워크숍 및 공모전에서 발표된 표절 문장 탐색 기법[18]과 본 논문에서 제안하는 기법과 비교하였으며 그 결과는 Table 1과 같다.

Table 1 Result Comparison with Sentence Plagiarism Detection method announced by PAN 2013 Workshop

Team	plagdet	Relative Improvement
R. Torrejón	0.82220	(-) 10.6
Kong	0.81896	(-) 10.2
Suchomel	0.74482	(-) 1.3
Our Method	**0.73505**	
Saremi	0.69913	(+) 5.1
Shrestha	0.69551	(+) 5.7
Palkovskii	0.61523	(+) 19.5
Nourian	0.57716	(+) 27.4
Baseline	0.42191	(+) 74.2
Gillam	0.40059	(+) 83.5
Jayapal	0.27081	(+) 171.4

Table 1에서 Team은 표절 문장 탐색 기법을 제안한 저자의 이름이며 그 중에서 Baseline은 PAN에서 기본적으로 제공하는 기법 그리고 Our Method는 본 논문에서 제안하는 기법을 나타낸다. 또한, 아래의 식을 사용하여 상대적 성능 향

상 지표(Relative Improvement)를 계산하였다.

$$\left(\frac{Our\,Method - Other\,Method}{Other\,Method}\right) \cdot 100 \qquad (9)$$

Table 1에서와 같이 본 논문에서 제안하는 워드넷 유사도 기반 신호화와 이산 푸리에 변환을 이용한 표절 문장 탐색(Our Method)이 4번째로 높은 성능을 보임을 알 수 있다. 결론적으로 유의어를 사용한 표절 형태뿐만 아니라 번역 및 요약 표절도 탐색할 수 있음을 알 수 있다.

5. 결론 및 향후 연구 방향

본 논문에서는 최초로 텍스트 데이터 마이닝에 신호 처리 분야에서 가장 중요한 이론인 푸리에 변환을 적용하였다. 적용 가능성을 확인하기 위해 다양한 실험을 한 결과, 텍스트 데이터 마이닝에 있어 이산 푸리에 변환은 유사한 패턴을 탐색할 수 있음을 확인하였고 명확한 성능 입증을 위해 표절 문장 탐색에 응용하였다.

이산 푸리에 변환을 적용한 표절 문장 탐색에서 우선적으로 문장을 신호화 하고, 상호 상관(Cross-Correlation)을 이용하여 표절 문장 탐색을 시도하였다. 실험 결과, 유의어를 사용하거나 번역 및 요약의 표절 형태는 탐색하지 못함으로써 높은 성능을 보이지 못했다. 이를 해결하기 위해 본 논문에서는 워드넷 유사도 기법을 이용하여 새롭게 문장을 신호화하고 이산 푸리에 변환을 적용하여 표절 문장을 탐색하였다. 실험 결과, 상호 상관을 이용하였을 때 보다 높은 성능 향상을 보였다.

본 논문에서 제안한 기법이 현재까지 제안된 표절 탐색 기법 보다는 월등히 높은 성능을 보이지는 않지만 신호 및 영상 처리에서만 사용되어 왔던 이산 푸리에 변환이 텍스트 데이터 마이닝에도 적용할 수 있음을 실험을 통해 증명 하였다. 따라서 현재의 연구를 기반으로 다양한 텍스트 신호화 기법과 이산 푸리에 변환을 적용 후 특징 추출 및 유사도 측정 기법에 대한 심화 연구가 진행된다면 국내뿐만 아니라 세계 최초로 이산 푸리에 변환을 적용한 텍스트 데이터 마이닝 연구 분야를 확립할 수 있다고 본다. 또한, 최근 딥러닝 기반으로 용어들의 의미적인 유사도를 계산할 수 있는 방법(Word Embeddings)이 제안되었는데 이러한 기술을 이용하면 계산 속도 및 성능 향상에 도움이 될 것이라고 기대한다.

References

[1] Cetin, E., Morling, R. C., and Kale, I. “An Integrated 256-Point Complex FFT Processor for Real-Time Spectrum Analysis and Measurement,” Instrumentation and Measurement Technology Conference, pp. 96-101, 1997.

[2] Briggs, W. L. and Henson, V. E, The DFT: an Owners′ Manual for the Discrete Fourier Transform, Society for Industrial and Applied Mathematics, 1995.

[3] Howell, K. B., Principles of Fourier Analysis, CRC Press, 2001.

[4] Lynn, P. A. and Fuerst, W., Introductory Digital Signal Processing with Computer Applications, John Wiley, 1998.

[5] Lee, C. H., “A Pattern Matching Algorithm using Correlation in Fourier Domain,” Journal of Korea Multimedia Society, Vol. 7, No. 9, pp. 1255-1262, 2004.

[6] Han, J. Y., Cho, C. H., and Son, I. S., “An Empirical Study on Corporate use of Big Data : The Case of Integrated Customer Log System at a Korean Home Shopping Firm,” Journal of Internet Electronic Commerce Research, Vol. 15, No. 6, pp. 1-19, 2015.

[7] Hwang, I. S., “A Study on Plagiarism Detection and Document Classification using

Association Analysis," Journal of Information Systems, Vol. 23, No. 3, pp. 127-142, 2014.

[8] Lyon, C., Malcolm, J., and Dickerson, B., "Detecting Short Passages of Similar Text in Large Document Collections," International Conference on Empirical Methods in Natural Language Processing, pp. 118-125, 2001.

[9] Lewis, J. P., "Fast Template Matching," Vision Interface, Vol. 95, No. 120123, pp. 15-19, 1995.

[10] Smith, J. O., Mathematics of the Discrete Fourier Transform (DFT): with Audio Applications, Julius Smith, 2007.

[11] Miller, G. A., "WordNet: A Lexical Database for English," Communications of the ACM, Vol. 38, No. 11, pp. 39-41, 1995.

[12] Wu, Z. and Palmer, M., "Verbs Semantics and Lexical Selection," 32nd Annual Meeting on Association for Computational Linguistics, pp. 133-138, 1994.

[13] Resnik, P., "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," 14th International Joint Conference on Artificial Intelligence, 1995.

[14] Jiang, J. J. and Conrath, D. W., "Semantic Similarity based on Corpus Statistics and Lexical Taxonomy," International Conference Research on Computational Linguistics, 1997.

[15] Leacock, C. and Chodorow, M., "Combining Local Context and Wordnet Similarity for Word Sense Identification," WordNet: An Electronic Lexical Database, Vol. 49, No. 2, pp, 265-283, 1998.

[16] D., "An Information-Theoretic Definition of Similarity," International Conference on Machine Learning, Vol. 98, pp. 296-304, 1998.

[17] Banerjee, S. and Pedersen, T., "An Adapted Lesk Algorithm for Word Sense Disambiguation using WordNet," International Conference on Intelligent Text Processing and Computational Linguistics, pp. 136-145, 2002.

[18] Cheema, W. A., Najib, F., Ahmed, S., Bukhari, S. H., Sittar, A., and Nawab, R. M. A, "A Corpus for Analyzing Text Reuse by People of Different Groups," 5th International Conference of the CLEF Initiative, 2014.

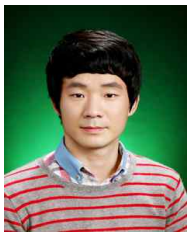
[19] Potthast, M., Hagen, M., Gollub, T., Tippmann, M., Kiesel, J., Rosso, P., and Stein, B., "Overview of the 5th International Competition on Plagiarism Detection," Conference on Multilingual and Multimodal Information Access Evaluation, pp. 301-331, 2013.

[20] Potthast, M., Stein, B., Barron-Cedeno, A., and Rosso, P., "An Evaluation Framework for Plagiarism Detection," 23rd International Conference on Computational Linguistics, pp. 997-1005, 2010.

[21] Lee, J. K. and Kim K. J., "Educational Contents and Implementation Procedures of the Training System for Research Ethics", Journal of the Korea Industrial Information Systems Research, Vol. 15, No. 5, pp. 235-246, 2010.

[22] Lee, J. S. and Park S. C., "The Document Clustering using Multi-Objective Genetic Algorithms", Journal of the Korea Industrial Information Systems Research, Vol. 17, No. 2, pp. 57-64, 2012.

[23] Choi, L. C., Park S. C., and Song, W., "Comparison of Document Clustering algorithm using Genetic Algorithms by Individual Structures", Journal of the Korea Industrial Information Systems Research, Vol. 16, No. 3, pp. 47-56, 2011.



이 정 송 (Lee Jung-Song)

- 정회원
- 전북대학교 전자정보공학부 공학사
- 전북대학교 전자정보공학부 공학석사

- 전북대학교 전자정보공학부 박사과정
- 관심분야 : 인공지능, 정보검색, 텍스트 데이터 마이닝, 디지털 아카이브즈

박 순 철 (Park Soon-Cheol)

- 종신회원
- 인하대학교 공과대학 공학사
- 미국 루이지아나 주립대학 전산학박사

- 한국전자통신연구원 근무
- 전북대학교 컴퓨터공학부 교수
- 관심분야 : 인공지능, 정보검색, 텍스트 데이터 마이닝, 디지털 아카이브즈