2007년도 대한전기학회 하계학술대회 논문집 2007. 7. 18 - 20

특정 조사와 빈도수 높은 단어를 이용한 한글 논문의 유사도 측정 시스템 구현

유승희*, 한소희, 조동섭
이화여자대학교 컴퓨터정보통신학과

Similarity Measurement System of Korean Documents Using the Specified Particles and High Frequency Words

Seung-hee Yoo*, So-hee Han, Dong-sub Cho
*Dept. of Computer science and En gineering, Ewha Womans Univ.

Abstract - 인터넷의 발달로 대량의 전자문서들을 손쉽게 구할 수 있는 정보의 바다라 불리는 현대사회에서 논문 표절은 심각한 문제를 안게 되었다. 표절여부를 검사하는 방법에는 여러 가지가 있지만 보다 정확하고 빠르게 검출할 수 있는 기법이 요구된다. 외국에서는 표절을 검사하기 위한 시스템적인 접근이 이루어지고 있지만 국내에서의 표절 검사에 대한 연구는 아직 초기 단계에 있다. 본 논문에서는 논문 표절 검사 시스템에 사용되는 기법 중 지문법을 바탕으로 하지만 기존의 단어, 문장 등을 사용하는 방법과 차별을 두어 몇몇 주요 단어와 특정 조사의 비교를 이용해 유사성을 측정하여 보다 빠르고 정확하게 검출할 수 있는 시스템을 구현해 보았다.

1. 서 론

논문을 쓸 때 타인이 발표한 데이터나 연구내용을 원본 자료의 출처를 밝히지 않고 자신의 글에 포함시키거나 출처를 밝혔다 하더라도 원문의 문장을 그대로 사용하는 것을 표절[標竊, Plagiarism]이라고 한다[1]. 전자문서 및 각종 매체의 발달로 데이터 수집이 용이해 짐에 따라 표절은 더욱 빈번해 졌지만 국내에서의 표절 연구는 아직 많이 미흡한 단계에 머무르고 있다[2].

본 논문에서는 한글 논문간 표절 여부를 검사하기 위한 시스템을 구현하고 있다. 본 시스템은 기존의 지문법을 사용하지만 기존 연구의 어절 별로 단어를 구분하고 그 단어들과 몇몇의 문장들을 서로 비교하는 방법과는 달리 5개의 빈도수 높은 단어와 조사를 사용하여 표절검사를 하는 방법을 사용하여 기존의 기법보다 정확도를 높이고 실행시간을 줄일 수 있고자 한다.

2. 본 론

2.1 기존의 표절 검사 기법

기존의 표절 검사 기법으로는 대표적으로 다음과 같은 2가지가 있다.

2.1.1 지문법(Fingerprint Method)

이 기법은 문장의 순서와 관계없이 사용된 단어의 빈도수, 문장, 문단의 평균길이, 특수문자의 사용횟수 등을 비교하여 표절 검사를 수행하는 방법으로서 온라인 표절 검사 사이트 Plagiarism.org[3], EVE2[4], Virmingham 대학의 CopyCatch[5], 전자 도서관에서 사용되는 COPS[6], WordCheck Ketword Software[7] 등의 표절 검사 시스템이 지문법을 사용하고 있다. 이 방법은 문서의 길이에 영향을 받지 않고, 지문을 뽑아내기만 하면 이 지문들끼리 비교를 통해 표절여부를 판단하기 때문에 시간 복잡도도 낮다. 하지만 단지 단어 등 지문의 빈도수 등을 이용하여 검사하기 때문에 단락, 혹은 문장의 구조적인 분석은 어려워진다. 구조적인 특징을 가지는 프로그램 코드 등의 검사에는 부적절한 방법이다.

2.1.2 구조기반 검사방법(Structure related Method)

이 기법은 일반 문서의 표절보다도 제어흐름을 가지고 있어 구조적인 특징을 가지고 있는 프로그램 소스 코드의 표절 검사에 주로 사용되는 기법으로 Plague[8], MOSS[9], CHECK[10] 등의 표절검사 시스템에서 사용된다. 프로그램 소스코드는 일반 문서와는 다리 변수 등을 다른 단어로 대체하는 표절하는 것이 가능하며 함수 등의 위치를 통체로 바꾸어 표절할 수 있으므로, 이러한 소스코드의 구조적 특징을 살펴보아 구조 기반 검사기법을 사용하는 것이 더 효과적이라 볼 수 있다.

<표 1> 지문법과 구조 기반 검사 방법의 비교

	지문법	구조기반검사기법
시간복잡도	낮음	높음
문서의 길이	영향 받지않음	길이에 따라 얻어내는 정보가 비례함
부분적인 표절탐지	어려움	쉬움

2.2 특정 조사와 빈도수 높은 단어를 이용한 유사도 검사

위의 표절검사 기법에서 보았듯이 문서의 표절검사에는 지문법이 더 적당하다고 볼 수 있다. 하지만 지문법은 문서의 길이에 영향을 받지 않지만 문서가 길어짐에 따라 검사를 해야 하는 지문(단어, 문장 등)의 수가 많아지기 때문에 속도가 느려지는 문제점이 있다. 또한 표절 논문이 아니더라도 비슷한 주제를 갖는 논문이라면 같은 단어를 사용하는 논문은 다수 존재한다. 그리고 표절 논문이라 하더라도 단어의 유사어 사용이 가능하기 때문에 검출이 어려운 경우도 있다. 그러므로 기존의 방법과 같이 단어나 문장 등만을 지문으로 삼은 방법은 정확도가 떨어질 수 있다.

이에 대한 개선 방법으로 본 논문에서는 관사, 접속사, 전치사 따위의 어미, 어형의 변화가 없는 품사(불변화사)가 오히려 표절 시 바꾸어 사용하기 어렵다는 특성을 이용하여 변형이 어려운 특정한 조사, 전치사와 빈도수 높은 단어 5개를 추출하여 두 논문의 유사도를 측정함으로서 표절 논문을 검사하였다. 이 방법을 사용함으로써 기존 방법보다 유사도를 높이고 실행 속도를 줄일 수 있을 것이라 예상된다.

2.3 특정 조사와 빈도수 높은 단어를 이용하 유사도 검사 방법

특정 조사를 비교한 유사도 검사방법은 크게 세 가지 단계로 나뉜다.

첫 번째 단계는 주요단어를 추출(Keyword Extraction)하는 단계로서, 표절논문이라면 빈도수 높은 단어 사용 횟수는 비슷할 것이라고 가정하여 A, B 논문 중 A 논문의 빈도 높은 단어를 추출한다.

두 번째 단계는 특정 조사의 추출(Particle Extraction)하는 단계로서, 기존의 단어, 문장 등만 추출하는 방법과 차별화하여 변형이 어려울 것이라 판단되어지는 영어의 전치사에 해당하는 한글의 조사와 그 빈도수를 추출한다. 추출조사에는 '위에', '안에', '부터', '까지', '에서', '보다', '하는', '위해', '되는', '에는', '하여', '간에', '의', '와', '과', '에', '은', '는' .등이 있다.

세 번째 단계는 앞에서 추출한 단어와 전치사를 통하여 그 유사도를 측정(Similarity Measurement)하는 단계로서, 일반적으로 텍스트 문서의 유사도를 측정하기 위해 일반적으로 VSM(Vector Space Model)을 사용한다. VSM[11]이란 문서들이 포함하는 단어들을 통해 그 문서를 재표현하는 하나의 방법으로 문서와 문서간의 유사성을 측정하여 그 문서를 분류한다. 이 방법에는 대표적으로 Cosine Similarity 계산법과 Euclidean distance 계산법이 있다.

본 시스템에서는 Cosine Angle을 통해 두 벡터가 얼만큼 유사한지 수치로 나타낼 수 있는 Cosine Similarity 계산법을 사용하였다. 그 식은 다음과 같다.

$$cosine\ \ similarity = \frac{A \cdot B}{\|A\| \|B\|} = \frac{a_1 b_1 + a_2 b_2 + \cdots + a_n b_n}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}\sqrt{b_1^2 + b_2^2 + \cdots + b_n^2}}$$

여기까지의 방법을 예를 들어 설명하면 다음과 같다.

나는 책과 볼펜을 사야한다 - 문서 1
나는 책과 샤프를 사야한다 - 문서 2

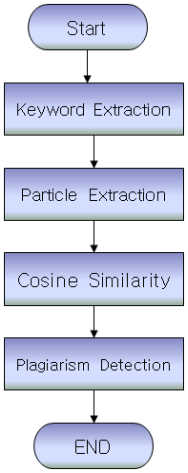
[나, 책, 볼펜, 샤프] - [단계 1]Keyword Extraction

[과, 을] - [단계 2] Particle Extraction

[나, 책, 볼펜, 샤프, 과, 을] - [단계 3] Similarity Measurement
A = [1, 1, 1, 0, 1, 1]
B = [1, 1, 0, 1, 1, 1]

이것을 풀면 유사도는 약 0.88로써 높은 유사도를 가지고 유사도가 높다는 것은 표절 가능성이 크다는 것을 나타낸다.

위의 과정을 순서도로 나타내면 다음과 같다.



<그림 1> 검사단계 순서도

다음은 본 시스템의 Pseudo code이다.

```
DATE structure :
    N is structure to save 명사s and it's count
    P is structure to save 전치사s and it's count


표절알고리즘(document[0..NUM_OF_DOCUMENT])

    ALLOCATE N[NUM_OF_DOCUMENT] and P[NUM_OF_DOCUMENT]
    INITIALIZE N[] and P[]

    OPEN document[0] and document[1] file

    WHILE index=0 to NUM_OF_DOCUMENT
        COUNT number of 명사 of document[index] and Save it to P[index]
        COUNT number of 전치사 of document[index] and Save it to N[index]
        SORT N[] and P[] by count
    END WHILE

    WHILE i=0 to NUM_OF_DOCUMENT
        WHILE j=i to NUM_OF_DOCUMENT
            CALCULATE similarity(P[i] and P[j], N[j] and N[j])
        END WHILE
    END WHILE

  RETURN SIMILAR Documents
END
```

<참조 1> 한글 논문 유사도 측정 시스템 Pseudo code

3. 결 론

본 논문에서는 빈도수 높은 단어 5개와 특정 조사를 이용하여 두 한글 논문의 유사도를 측정하여 표절 검사를 하는 프로그램을 구현해 보았다. 이 기법은 다음과 같은 기대를 할 수 있다.

먼저, 비교 단어의 수를 대폭 줄임으로써 실행 속도를 줄일 수 있다. 그리고 어형, 어미의 변화가 없는 조사를 비교함으로써 유사단어 사용 시 표절 추출이 어려운 단어 검사보다 표절 검사의 정확도를 높일 수 있다.

본 논문에서는 표절 검사 프로그램을 구현해 보았으며 향후에는 실험을 통해 얻어지는 통계로 본 시스템의 정확도와 속도를 비교, 검증할 필요가 있고 시스템을 가시화하여 사용 시 편리하게 구현 할 필요가 있다.

[참 고 문 헌]

[1] http://www.calstatela.edu/centers/write-on/plagiarism.h
[2] http://www.gyosuclub.co
[3] http://www.plagiarism.org
[4] http://www.canexus.com/eve/abouteve.html
[5] http://www.copycatch.freeserve.co.uk
[6] Sergey B, James D, and H .G, "Copy detection mechanisms for digital documents", *Proc. ACM SIGMOD Internagional conference on Management of date,* pp. 398-409, 1995
[7] http://www.wordcheksystems.com/
[8] Antonio.S., Hong V.L., and Rynson. W.H.L., "CHECK : A document plagiarism detection system," *Proc. ACM Symposium on Applied Computing,* pp. 70-77, 1997
[9] Whale, "Identification of Program Similarity in Large populations," *The Computer Jornal,* Vol.33, No.2, pp. 140-146
[10] Kevin W. Bowyer and Lawrence O. Hall, "Experience Using "MOSS" to Detect Cheating On Programming Assignments" *Department of Computer Science and Engineering University of South Florida*
[11] Lee, D.L.; Huei Chuang; Seamons, K. "Document ranking and the vector-space mode"l Software, IEEE