주변 문장 유사도를 이용한 문서 재사용 측정 모델

최성원 김상범 임해창
고려대학교 컴퓨터학과 자연어처리 연구실
{swchoi, sbkim, rim}@nlp.korea.ac.kr

A Text Reuse Measuring Model Using Circumference Sentence Similarity

Sungwon Choi Sangbum Kim Haechang Rim
Dept. of Computer Science, Korea University, Seoul, Korea

요 약

기존의 문서 재사용 탐지 모델은 문서 혹은 문장 단위로 그 내부의 단어 혹은 n-gram을 비교를 통해 문장의 재사용을 판별하였다. 그렇지만 문서 단위의 재사용 검사는 다른 문서의 일부분을 재사용하는 경우에 대해서는 문서 내에 문서 재사용이 이루어지지 않은 부분에 의해서 그 재사용 측정값이 낮아지게 되어 오류가 발생할 수 있는 가능성이 높아진다. 반면에 문장 단위의 문서 재사용 검사는 비교 문서 내의 문장들에 대한 비교를 수행하게 되므로, 문서의 일부분에 대해 재사용을 수행한 경우에도 그 재사용된 부분 내의 문장들에 대한 비교를 수행하는 것이므로 문서 단위의 재사용에 비해 그런 경우에 더 견고하게 작동된다. 그렇지만, 문장 단위의 비교는 문서에 비해 짧은 문장을 단위로 하기 때문에 그 신뢰도에 문제가 발생하게 된다. 본 논문에서는 이런 문장단위 비교의 단점을 보완하기 위해 문장 단위의 문서 재사용 검사를 수행 후, 문장의 주변 문장의 재사용 검사 결과를 이용하여 문장 단위 재사용 검사에서 일어나는 오류를 감소시키고자 하였다.

1. 서론

문서 재사용은 "다른 저자의 문서를 변형하여 새로운 문서를 만들어내는 작업"을 통칭한다. 문서 재사용의 문제는 과거부터 학생들의 리포트나 논문의 작성시 다른 책의 내용을 재사용 하는 정도의 형태로 존재해 왔지만, 최근의 인터넷, 검색엔진의 발전과 워드프로세서의 사용증가로 인해 재사용 대상 문서를 쉽게 찾을 수 있고 그 편집 또한 쉬워짐에 따라 학생들의 리포트, 논문, 신문 기사 등에서 이런 문서 재사용이 광범위하게 이루어짐으로 해서 문서 재사용의 탐지에 대한 연구의 필요성이 대두되었다.

문서 재사용 탐지를 성공적으로 수행하기 위해서는 다음 두가지 조건을 만족해야 한다.

첫째는 당연한 이야기 이지만, 문서 재사용이 이루어진 부분을 정확하게 찾아내야 한다. 문서 재사용은 기존의 문서를 그대로 사용하기보다는 문서를 재사용하였다는 의혹을 피하기 위해 단어의 삽입, 삭제, 어순의 변경, 문장 구조 변경 등 문서의 변형을 가하여 이루어지고 특히 대상문서를 한국어로 했을 때, 다른 언어와는 다른 한국어의 자유 어순이라는 특성이 문서의 재사용을 더욱 다양하고 복잡하게 할 수 있음을 고려해야 한다. 또한, 재사용의 대상 문서로써 문서 하나를 통째로 재사용하기 보다는 필요에 따라 부분을 발췌하여 재사용하는 경우가 더 빈번히 이루어 지므로, 그에 알맞게 재사용 된 부분을 정확히 찾아낼 수 있는 검사를 수행해야 한다.

둘째, 효율적인 검사를 수행해야 한다.문서의 재사용 여부를 정확히 검사해 내기 위해서는 검사를 수행하는 대상 문서 집합 내에서 문서 재사용이 이루어 졌는지를 검사하는 것도 중요하지만, 그보다 먼저 실제 재사용의 출처가 되는 문서가 검사 대상 문서 집합 내에 존재해야 재사용 여부에 대한 판단이 가능하다. 때문에 재사용 탐지 시스템은 웹 문서를 비롯하여 다른 논문/리포트/신문기사 등 재사용의 출처 문서의 가능성이 있는 문서를 최대한 포함하여 검사를 수행해야 한다. 그를 위해 문서 재사용 검사 시스템은 검사 대상 문서 집합 내의 문서에 대한 문서 재사용 검사를 사용자가 허용할 수 있는 합리적인 시간과 자원의 한도내에서 검사할 수 있는 효율성을 가져야 한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 문서 재사용에 대한 국내외의 연구를 살펴보고 기존 연구의 문제점을 해결하기 위한 본 논문에서의 연구 동기를 밝히고, 3장에서 제안하는 문서 재사용 검사 모델에 대해 살펴보고, 4장에서 제안하는 시스템의 성능에 대해 살펴보고 더 나아가 기존의 시스템과의 비교를 수행하고, 마지막 5장에서 결론을 맺도록 하겠다.

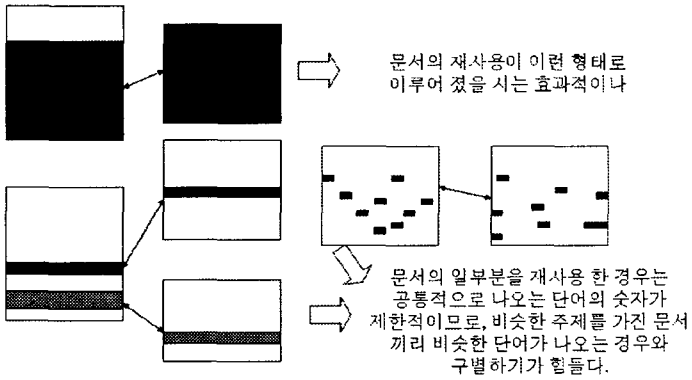
2. 관련 연구 및 연구동기

2.1 문서 단위 문서 재사용 검사 연구

문서 재사용을 판별하는데 있어서 비교하고자 하는 두 문서에서 사용된 단어 혹은 n-gram을 추출하여 문서를 비교하는 방법이다. 이에 속하는 기존연구로는 문서 간의 n-gram 중복을 측정하는 지문법[2]과, 정보 검색 모델을 이용하는 방법[1,6] 등이 있다.

지문법은 두 문서 내에 존재하는 n-gram을 추출하고 그 중에 얼마나 많은 n-gram이 공통적으로 존재하는 가를 문서 재사용의 판단 기준으로 삼는 방법이며, 정보 검색 모델을 이용한 연구는 문서에서 색인어를 추출하여, 문서와 문서간의 재사용 검사를 비교하고자 하는 두 문서 중 한 문서를 질의어로 사용하여 문서의 재사용에 대한 검사를 문서간의 유사도를 측정하는 개념으로 치환하여 문제를 해결하고자 하는 방법이다.

이런 문서 단위의 재사용 검사는 [그림 1]에서 볼 수 있듯이, 문서 전체를 재사용 한 경우에는 좋은 성능을 보일 수 있으나 여러개의 문서에서 일부분을 발췌하여 재사용한 경우에는 두 문서 사이에 공통적으로 나타난 부분이 적게 되므로 주제에 따라 공통적으로 사용될 단어나 n-gram이 유사한 주제에 대해서 정확한 분류를 수행하는 것이 어렵게 되는 단점이 있다.



[그림 1] 문서 단위로 문서 재사용 검사를 했을 시 발생할 수 있는 문제

때문에 부분적인 문서 재사용을 판정하기 위해서는 문서를 보다 작은 단위로 잘라서 문서 재사용을 판정하는 것이 필요하게 된다.

2.2 문장 단위의 문서 재사용 검사

문서의 문서 재사용을 검사하는 데 있어서, 문서를 문장을 단위로 쪼개고 그 문장들과 다른 문서 내의 문장에 대한 비교를 수행하여 두 문서내에 문서 재사용이 이루어진 문장이 얼마나 많은가를 재사용 판단의 척도로 사용하는 연구이다. 문장과 문장을 비교하는 방법으로는 단순히 같은 단어가 사용되었는지를 가지고 재사용을 판정하는 지문법을 쓰기보다는 보다 정확한 문장간의 비교를 위해 문자열 비교 알고리즘[1,3] 등을 사용한다.

문자열 비교를 이용한 문서 재사용 검사는 기존의 영어권에서 만들어진 문서 재사용 검사 시스템에서 널리 쓰이는 방법[1]으로, 비교하고자 하는 두 문서 내의 모든 문장과 문장 간에 문장의 변형을 고려하여 두 문장 간의 차이를 허용하는 문자열 비교를 통해서 두 문장 내에 존재하는 일정 길이 이상(최소 매칭 길이)의 공통 문자열을 추출하고 그것이 문장 내에서 얼마나 많은 비율을 차지하는 가를 문장간의 재사용 여부를 판단하는 척도로 삼는다.

문장 단위의 문서 재사용 검사는 문서를 문장 단위로 쪼개서 문서 재사용을 검사하므로, 두 문서간에 일부분만이 재사용 관계에 있을 때에도 견고하게 작동한다.

그러나 효율성의 관점에서 문장 단위의 문서 재사용 검사를 살펴보면, 문서 집합이 n개의 문장으로 이루어져 있을 시, 약 n x (n - 자기 문장의 문장수) 번의 검사를 수행해야 하므로, 문서 집합이 증가함에 따라 수행량이 급격하게 증가하는 경향을 보인다. 예를 들어, 평균 문장수 200 문장으로 이루어진 학생들의 레포트 50개와 그 출처 후보 문서 50개에 대해 문서 재사용 여부를 검사하려 한다면, 약 3억9천6백만번의 문장 비교를 수행해야 한다. 문자열 비교 알고리즘을 사용하는 경우, 그 알고리즘은 $O(n^2)$~$O(n^4)$의 시간 복잡도를 가지게 되므로, 그 검사의 수행량은 급격하게 증가하게 된다. 때문에 효율성 증대를 위해 문자열 검색 이전에 문서 재사용 가능성이 어느정도 있는 문장쌍을 추출하여 불필요한 검사를 줄이는 과정이 필요하다.

검사 정확도의 측면에서 문장 단위의 문서 재사용 검사를 살펴보면 문장 단위로 검사를 수행하기 때문에 같은 속담, 명언, 관용어구를 사용하거나 우연에 의해서 문장 내에 사용한 단어가 비슷해지는 경우 등에 의해 문장 간의 재사용 판단에 오류가 생길 가능성이 생긴다. 이런 현상은 문장이 짧아지면 짧아질 수록 더욱 빈번히 발생한다. 다음 예의 두 문장 A,B에 대해서,

A : 둘다 문서 작성에 많이 사용되고 있다.
B : PDF 어도비사의 아크로뱃에서 사용되는 문서 작성용 파일 포맷으로 파일 용량이 작고, 강력한 특성 때문에 많이 사용되고 있습니다.

A, B간에 문자열 비교를 수행하면 공통 문자열로 '문서 작성', '많이 사용되고 있다.'가 추출되고, A 문장의 5/6이 공통적으로 사용되었기 때문에, A문장이 B문장을 재사용 하였다고 판정하는 오류를 범하게 된다. 위의 문장의 경우와 같이 한개의 문장 만으로는 다른 문장과 공통적인 문자열이 큰 부분을 차지한다 할지라도, 재사용으로 판정할 수 없는 경우가 있기 때문에, 보다 정확한 문서 재사용을 판정하기 위해서는 비교 문장의 주변의 문장을 살펴봐야 할 필요성이 있다.

또한 문장의 재사용을 검사시의 최소 매칭 길이의 결정에 있어서, 최소 매칭 길이를 작게 설정해 주면 문장의

변형이 많이 이루어진 경우에도 재사용 된 문장을 찾을 수 있으나, 그 판단의 확실성은 떨어지게 된다. 반면에 최소 매칭 길이를 크게 설정해 주었을 경우에는 두 문장 간의 재사용 여부 판정의 확실성은 보장할 수 있으나, 문장의 변형에 대해서는 견고하지 못하게 된다. 간단한 예로써, 다음 세 문장을 보도록 하자.

A: MPEG2는 현재 DVD등의 컴퓨터 멀티 미디어 서비스, 직접 위성방송, 유선방송, 고화질 TV등의 방송 서비스, 영화나 광고 편집등에 널리 사용되고 있습니다.
B: MPEG2는 DVD, 위성방송, 고화질 TV, 유선방송 등과 같은 영화,방송, 광고 분야에 널리 쓰이고 있습니다.
C: 사용자는 고화질 TV를 통해 각종 DVD 미디어, 위성방송, 유선방송 등에서 제공하는 서비스를 좀 더 좋은 품질로 즐길 수 있다.

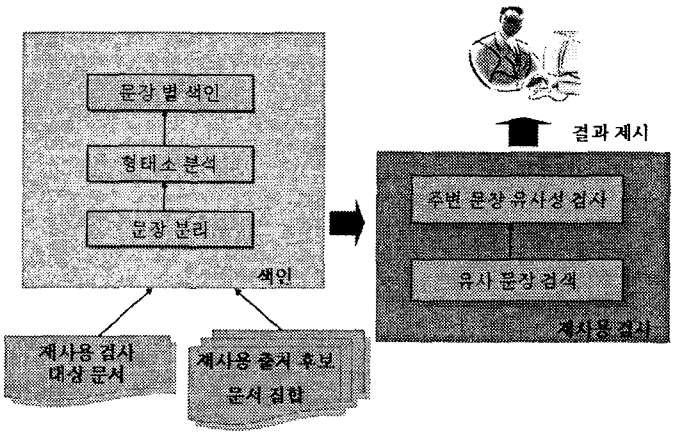
A문장과 B문장의 관계는 단어의 나열 위치의 변경과 단어의 삽입,삭제,변경를 통해 문서를 재사용 했다고 볼 수 있는 문장이고, A문장과 C문장은 서로 다른 내용을 담고 있는 문장이다. 만일 최소 매칭 길이를 1로 짧게 설정하였을 시는, A와 B간의 매칭 문자열은 {MPEG2, DVD, 위성방송, 고화질 TV, 유선방송, 영화, 방송, 광고, 널리, 있습니다}의 문자열이 매칭 될 것이다. 두 문장간에 매칭된 문자열이 상당 부분 존재하기 때문에 두 문장은 문서 재사용이 이루어 진 문장으로 추정할 수 있게 된다. 그렇지만 최소 매칭 길이를 작게 주었기 때문에, A문장과 C문장 간에도 {고화질 TV, DVD, 미디어, 위성방송, 유선방송, 서비스}의 전체 문자열 중 상당 부분이 매칭 되어 재사용 된 문장으로 판정하는 오류를 범하게 된다. 반면에 최소 매칭 길이를 늘리면 B,C 문장 모두를 문서 재사용이 이루어지지 않은 문장으로 판정하는 오류를 범하게 된다.

2.3 연구 동기

앞에서 살펴본 기존의 문서 단위의 문서 재사용 검사에 대한 연구는 그 단위가 너무 크기 때문에 문서의 부분적인 재사용에 대한 판정이 어려운 점이 존재하고, 문장을 단위로 한 경우 문서 재사용의 판정 기준을 느슨하게 주었을 시에는 문장의 변형에 대해서도 견고하게 작동하나, 그 확실성이 떨어지는 단점이 있으며, 판정 기준을 엄격하게 두었을 시에는 문서의 변형에 대해서 견고하지 못하게 되는 단점이 있다. 이런 단점을 보완하기 위해 본 논문에서는 문서의 부분적인 재사용에 대한 검사를 견고하게 수행하기 위해, 문장 단위로 문서 재사용 검사를 수행하도록 하며, 문장 단위의 문서 재사용 검사를 문장 내의 단어를 기반으로 유사 문장을 검색하는 관점에서 수행하여 검사의 효율성을 향상시키고자 한다. 그리고 위의 문장 A,B,C의 예와 유사한 문제를 해결하기 위해 문장의 재사용 판단의 기준은 느슨하게 두고 문서의 재사용은 한 문장이 아닌 여러 문장을 단위로 나타난다는 특성을 이용하여 비교하고자 하는 문장의 앞 뒤 문장의 비교를 통해 발생가능한 오류를 줄일 수 있는 모델을 제안하고자 한다.

3. 제안하는 방법

제안하는 문서 재사용 검사 방법은 크게, 1) 문서 집합에 대한 색인 단계와 2) 색인된 내용을 가지고 실제 재사용이 이루어 졌는지를 검사하는 재사용 검사 단계의 두 단계로 나눌 수 있다. 제안하고자 하는 모델을 이용한 시스템의 구성도는 다음 [그림2]과 같다.



[그림 2] 시스템 구성도

3-1. 색인 단계

색인 단계는 입력 받은 문서를 검색에 사용할 수 있도록 색인하는 작업을 칭한다. 색인 작업은 다음과 같은 단계로 이루어진다.

1) 문장 분리

문장 단위로 검사를 진행하므로 먼저 문서를 문장 단위로 분리해 주어야 한다.

2) 형태소 분석

전명재(2004)의 연구[7]에 의하면, 문장의 주요 형태소만으로 문서 재사용을 검사하는 것은 원문의 특성을 그대로 살릴 수 있으면서도, 비교량을 줄여주어 검사의 효율성을 증대시킬 수 있으며, 조사등을 배제함으로써 한국어의 자유로운 어형 변화를 이용한 재사용에도 견고하게 대응할 수 있게 해주기 때문에, 문장에 대해 형태소 분석을 수행하여 주요 형태소를 대상으로 색인어를 추출하도록 한다.

3) 문장 별 색인

추출된 형태소를 가지고 문장 별로 색인을 수행한다.

3-2 재사용 검사

색인된 결과를 가지고 문서 재사용 여부를 검사하는 작업이다. 이 작업은 다음과 같은 단계로 이루어진다.

1) 유사 문장 검색

문장을 검색하는 모델은 Vector-Space model에서 가중치 계산식을 okapi-model의 가중치 계산식으로 변경하여 사용하였다[4].

특정 단어 k의 I문장에서의 가중치 W_{ki}는 다음과 같이 계산하였다.

$$W_{ki} = \log\frac{N}{sf} \times \frac{(k_1+1)tf}{K+tf}$$

$$K = k_1((1-b)+b\frac{Sentence\ Length}{Average\ Sentence\ Length}$$

k_1 : 1.2, b : 0.75

N : 전체 문장 개수

계산된 색인어의 가중치를 이용하여 두 문장 S_i와 S_j간의 유사도 값을 다음과 같이 계산한다.

$$Sim(S_i, S_j) = \sum_{k=1}^{t} w_{ki} \times w_{kj}$$

계산된 유사도는 vector-space model의 가중치 식을 따르지 않았으므로 정규화 된 값을 가지지 않기 때문에, 그 문장의 길이가 길수록 높은 값을 가질 수 있게 된다. 때문에 유사도에 대한 정규화 작업이 필요하게 되는데 본 연구에서는 특정 문장 S_k에 대한 다른 문장과의 유사도 정규화를 자신과 동일한 문장이 가지게 되는 유사도 값 $Sim(S_k, S_k)$를 기준으로 하였다. 이를 이용한 두 문장의 재사용 가능성을 추정하는 함수 $Eval(S_a, S_b)$는 다음과 같다.

$$Eval(S_a, S_b) = \begin{cases} 1 & if, Sim(S_a,S_b) \geq \delta_1\ or\ Sim(S_b,S_a) \geq \delta_2 \\ 0 & otherwise \end{cases}$$

$$\delta_1 = Sim(S_a,S_a)\times\alpha, \delta_2 = Sim(S_b,S_b)\times\alpha$$

α값은 실험에 의해서 결정되며 단어의 삽입,삭제,교체등을 고려하여 재사용된 문장의 대부분을 걸러낼 수 있도록 설정한다. Sa에 대한 Sb의 재사용 가능성을 추정하는데 Sim(Sb,Sa)를 동시에 사용하는 것은 긴 문장을 작은 문장으로 쪼개어 문서 재사용을 수행한 경우, 긴 문장에 대해 쪼개진 문장들은 낮은 유사도 값을 가지게 되는 문제를 해결해 주기 위한 것이다.

2) 주변 문장 유사성 검사

1)의 단계에서 발생할 수 있는 재사용 판단 오류를 수정하기 위해 어떤 문장과 그 문장이 재사용한 것으로 추정되는 문장의 주변 문장도 역시 재사용으로 추정되었는가를 통해 재사용을 판단하도록 한다.

본 논문에서는 재사용으로 추정된 A문서의 i번째 문장 S_{Ai}와 B문서의 j번째 문장 S_{Bj}에 대한 재사용 판단 함수 $Det(S_{Ai}, S_{Bj})$를 다음과 같이 정의하였다. 여기서 k값은 주변 몇 개의 문장을 재사용 판단에 이용할 것인가를 뜻한다.

$$Det(S_{Ai}, S_{Bj}) = \sum_{a=-k}^{k}\sum_{b=-k}^{k} \lambda Eval(S_{Ai+a}, S_{Bj+b})$$

$$\begin{cases} \lambda = 0 & if\ ab < 0 \\ \lambda = 1 & if\ ab \geq 0 \end{cases}$$

여기서, λ값이 0과 1로 나뉘는 것은 주변 문장을 비교할 때 단순히 두 문장의 주변에 있는 모든 문장 간에 비교를 수행하기 보다는 앞에 존재하는 문장과 뒤에 존재하는 문장을 구분하여 비교하기 위한 것이다.

재사용 판단함수 $Det(S_{Ai}, S_{Bj})$이 일정 임계값 이상, 즉 주변 문장이 일정 숫자 이상만큼 재사용한 것으로 추정된다면, 그 문장을 재사용 문장으로 판정하였다.

전체 문장에 대한 문서 재사용 검사가 종료된 뒤에는 각 단일 문서의 재사용 정도를

$$문서\ 재사용\ 정도 = \frac{재사용\ 판정\ 문장\ 수}{전체\ 문장\ 수}$$

로 계산하여 사용자에게 제시하게 된다.

5. 실험 및 결과 분석

실험에서 사용되는 문서 집합은 실제 수업에서 제출된 리포트와 리포트가 문서 재사용을 수행한 출처 웹 문서로 구성된다. 문서 집합의 주제는 "IT 신기술에 관한 조사"이고, 다음과 같이 구성되어 있다.

리포트 집합 : 총 26개, 평균 문장 수 : 113개	
재사용 문서	20개
재사용 안한 문서	6개
웹 문서 집합 : 총 15개, 평균 문장 수 : 143개	
문서 집합 총 문장 수	4900개

본 논문에서 제시하는 모델은 '문서 재사용이 이루어진 부분을 정확하게 찾는 것'을 목적으로 하기 때문에 실험 결과는 '찾은 문장이 실제로 문서 재사용을 통해 생성된 문장인가', 그리고 '전체 문서 재사용을 통해 생성된 문장 중 얼마나 많은 문장을 찾았는가'를 가지고 평가하도록 하겠다.

실험 문서 집합의 재사용 정답 집합은 문서 재사용 관계를 가진 문장과 문장의 쌍으로 이루어져 있으며, 그 구축과정은 다음과 같다.

1) 본 논문에서 제안한 모델을 이용한 시스템으로 문서 재사용 검사 및 검사 결과의 검토 및 수정하여 1차 정답 집합 생성

2) 문자열 매칭 알고리즘을 이용한 시스템으로 문서 집합 검사 결과를 1차 정답 집합과 비교 후 차이가 나는 부분에 대한 재검토를 통해 2차 정답 집합 생성

3) 제안한 모델을 이용한 시스템과 문자열 매칭 알고리

즘을 이용한 시스템의 임계값을 낮춰가면서 기존의 정답 집합에서 추가적으로 재사용 판정을 받은 문장 쌍들에 대해 재사용 여부 검토 후 정답 집합에 추가.

검사 결과 검토 과정에서 두 문장간을 재사용 관계로 판정하는 기준은 '두 문장이 같은 내용을 담고 있을 때, 또는 두 문장의 내용이 포함관계에 있을 때'로 두었다.

문서 재사용 검사 시스템을 평가하기 위한 평가 지수는 다음과 같이 재사용 판정 문장 쌍에 대한 정확률과 재현률을 사용하였다

$$\text{정확률}(precision) = \frac{\text{시스템이 제시한 검사 결과 중 정답 집합에 속한 문장 쌍의 개수}}{\text{시스템이 제시한 검사 결과 문장 쌍의 개수}}$$

$$\text{재현률}(recall) = \frac{\text{시스템이 제시한 검사 결과 중 정답집합에 속한 문장쌍의 개수}}{\text{재사용 정답 집합의 문장쌍의 개수}}$$

유사 문장 검색 함수 Eval(S_a, S_b)에서 재사용 판단을 위한 유사 문장을 추정하는 임계값 α와 주변 문장 검사 시의 범위값 k, 그리고 주변 문장내의 유사 문장 개수를 변화 시키며 실험한 결과값은 다음과 같다

(P : 정확률, R : 재현률, 단위 %)

α값	0.4		0.5		0.6	
	P	R	P	R	P	R
k=0	49.7	99.0	71.3	97.1	94	93.7
k=2, 1	76.9	96.4	94.6	95.1	99.9	90.5
k=2, 2	86.2	91.6	96.7	89.8	99.9	84.5
k=2, 3	90.9	81.0	98.2	80.3	99.9	73.1
k=3, 1	74.8	96.7	93.7	95.1	99.7	92.1
k=3, 2	83.4	93.9	94.2	91.4	99.9	89.8
k=3, 3	86.9	93.2	96.9	85.2	99.9	81.9
k=3, 4	89.7	81.1	97.4	77.8	99.6	75.7

[표 1] 임계값의 변화에 따른 성능 변화

[표 1]을 분석해 보면 임계값 α가 내려감에 따라 재현률은 올라가고 정확도는 떨어지는 것을 볼 수 있다. 또한 검사하는 주변 문장의 크기가 커질수록 재현율이 올라가고 그 정확도는 약간 감소하는 것을 알 수 있다. 주변 문장 내에서 유사 문장의 개수의 임계값은 증가할수록 정확도는 증가하는데 비해, 재현율의 감소는 눈에 띈다.

검사 결과에서 대한 오류 분석 결과, 정의문을 비롯한 비교적 정형화된 정보를 제공해주는 문장을 주변 문장의 내용에 상관없이 같은 정보를 가진 문장으로 취급하여 정답 집합에 추가해 준 것들이 주변 문장을 사용한 모델의 재현율의 감소를 가져왔다.

기존에 제시되었던 문서 재사용 검사 방법인 문자열 비교 방법 중 Greedy String Tiling 알고리즘을 사용한 시스템과 비교한 결과는 다음과 같다.[1]

	Precision	Recall
α =0.6, k=3,2	99.7%	92.1%
문자열 비교	97.6%	90.5%

문자열 비교의 주 오류는 전의 기존 연구에 대한 내용에서 밝혔듯이, 문장이 짧아질 수록 오류를 많이 발생하는 경향을 가지게 된다. 특히 실험 문서 집합이 신기술에 대한 문서였으므로, 과학 기술과 관련된 조직명과 기술명의 약자와 풀네임이 병기되어 있는 형태가 많이 존재하였기 때문에 단지 그 조직명과 기술명이 문장내에 존재하는 것으로 문서 재사용이 이루어진 문장이라고 분류하는 오류가 발생하였다.

6. 결론

문서, 문장 단위의 문서 재사용 검사의 단점을 극복하기 위해, 본 연구에서는 문장 단위를 기반으로 그 주변 문장에 대한 비교까지로 모델을 확장하는 방법을 사용하였다. 실험 결과 주변 문장의 유사도를 비교하는 것이 성능 향상에 도움을 주었으며, 그 성능은 기존의 문장 단위 문서 재사용 검사 방법 에 비해 어느 정도의 향상을 가져왔으며, 검색 기반의 문장 검사를 수행함으로써 문서 재사용 검사의 효율성을 증대시킬 수 있었다.

7. 참고 문헌

[1] Paul D. Clough, "Measuring Text Reuse"

[2] Caroline Lyon, "Detecting short passages of similar text in large documents collection"

[3] Wise M, "Running Karp-Rabin Matching and Greedy String Tiling.'

[4] S E Robertson, "Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track"

[5] Narayanan Shivakumar, "SCAM : A Copy Detection <echanism for Digital Documents"

[6] 전명재, "한글 구조특성과 지역정렬 알고리즘을 사용한 표절 판정 시스템의 개발"

[7] 전명재, "개선된 Local Alignemtn와 단어 축약 기법을 이용한 한글 문서의 효과적인 표절 검사"