한글 구조특성과 지역정렬 알고리즘을 사용한 표절 판정 시스템의 개발

전명재[0+] 박상돈 박웅 허진영 조환규[+]
부산과학고등학교 부산대학교[+]
mjjun@pearl.cs.pusan.ac.kr johnsdpark@korea.com feelblu@korea.com
mathclear@hanmail.net hgcho@pusan.ac.kr

Plagiarised Reports Detection System using Characteristcs of Korean Language and Local alignment Algorithm

Jun M.J[0] Park S.D. Park W. Heo J.Y. Cho H.G.
Pusan Science Academy, Pusan National University

요 약

최근 논문의 표절 및 저작권과 관련하여 여러 가지 사건들이 일어나 많은 관심과 우려를 불러일으키고 있다. 특히 인터넷 통신의 발달 및 워드프로세서의 기능 향상으로 인해 일선 교육현장에서의 표절에 관한 문제는 더욱 커지고 있다. 하지만 문서의 표절 여부를 가려내는 작업은 쉬운 일이 아니다. 과제로 제출되는 일반 문서의 경우 본문의 내용이나 문서의 개수를 고려해 볼 때 사람이 직접 표절 여부를 검사하는 것은 매우 힘든 작업이다. 그리고 어간, 어미의 변형이 쉽게 일어날 수 있는 한글의 경우에는 영어에서처럼 어절 단위로 두 문서를 비교하여 표절여부를 판정하는 기존의 방법은 적합하지가 않다. 본 논문에서는 한글로 작성된 텍스트 문서의 표절 여부를 효과적으로 검출해 내기 위한 새로운 방법들을 제시하고 있다. 그리고 실제로 수집된 다양한 문서 데이터 집합들에 대해 각각의 방법들을 테스트해 보고 실제 데이터에서 가장 효율적인 방법이 어떤 것인지 제시한다.

1. 서 론

문서 관련 저작물의 표절 여부는 끊임없이 제기되는 문제이다. 크게는 권위있는 논문이나 기사에서부터, 작게는 학교의 과제물에 이르기까지 문서의 표절은 많은 문제를 야기하고 있으며 빨리 극복되어야 하는 문제 중 하나이다. 그러나 인터넷 발달로 인한 전자 문서의 양적 증가와 데이터 수집의 용이성으로 인해 문서의 표절 여부를 검사하는 것은 갈수록 어려운 일이 되고 있다. 특히 한글 텍스트 문서의 경우 조사나 어미의 변형이 쉽게 일어나고, 영어에 비해 글자의 수가 매우 많다는 것을 고려해 볼 때 영어 문서의 유사성을 측정하는 기존 방법들로 한글 문서의 표절 여부를 검사하는 것은 여러 가지 문제점을 가지고 있다. 본 논문에서는 간단한 문서의 카피 뿐만 아니라 악의적인 표절도 검사할 수 있는 강건한 한글 문서 표절 검사 시스템을 개발하기 위한 여러 가지 방법들을 제시하고, 실제로 수집한 여러 가지 데이터 집합에 대해 각 방법들을 적용해 보았다. 그리고 각 방법들의 성능을 테스트해 봄으로써 한글의 표절 여부를 가장 효과적으로 검사하는 데에 있어서 고려해야 할 사항들을 살펴보았다.

2. 한글 문서와 표절의 특징

문서의 표절 여부를 검사하는 기존의 방법은 주로 어절 별로 문서의 단어를 구분하고 이 단어들을 서로 비교하는 방법이다. 하지만 이와 같은 어절 단위로의 구분은 어절의 변화가 거의 없는 영어에는 매우 적합하지만 어간, 어미의 변화가 자유로운 한글에는 꼭 들어맞는 방법이 아니다. 그리고 이와 같이 문서 전체를 고려하는 방법은 문서의 길이가 늘어날 수록 효율이 떨어지기 때문에 과제물과 같이 비슷한 주제의 상당수의 문서에 대한 표절 검사를 수행할 경우 상당한 비용을 감수해야 한다. 더구나 문서의 편집, 재가공이 더욱 늘어나고 있는 상황에서 지능적인 표절에 대한 대응을 하기 위해서는 더 세련되고 효율적인 표절검사방법이 필요하다. 기존의 문서 유사성 검사에 비해 더 효율적으로 한글 문서의 표절여부를 검사하기 위해서는 다음과 같은 특징을 고려해야 한다. 첫째, 한글은 영어에 비해 글자수가 2byte로 표현되며 그만큼 캐릭터 수가 많다. 둘째, 문맥에 큰 영향이 없는 어간, 어미나 조사의 변화가 자유롭다. 셋째, 품사의 위치가 비교적 자유롭다. 보다 효율적인 한글 표절 검사 시스템을 위해서는 이런 사항들이 고려되어야 한다.

지능적인 일반적 표절을 살펴보면 표절의 유형은 위에서 언급한 한글 구조의 특징과 많은 연관이 있음을 알 수 있다. 요약해 보면 (a)문장의 위치 바꾸기 (b)문장의 짜깁기 (c)어간이나 어미의 변형 (d)부사의 추가나 위치 변경 (e)조사의 변경 (f)단어치환 등의 방법으로 지능적인 표절이 행해진다.

3. 표절문서 탐색 알고리즘 제안

3.1 지역 정렬 알고리즘을 사용한 유사도 측정

두 문서간의 유사성을 측정하기 위한 알고리즘으로는 지역정렬(local alignment) 알고리즘을 사용하였다. 정렬 알고리즘은 두 서열을 가장 잘 매치되게 정렬시켜주는 알고리즘으로서 지역 정렬은 두 서열의 가장 유사한 부분을 매칭시켜 주는 방법이다. 문장의 순서에 관계없이 특정 단어의 빈도수로 표절 검사를 수행하는 지문법 등과 달리 문장의 순서를 고려하여 표절 여부를 검사하고자 하는 경우에는 정렬과 같은 방법이 필요한데, 정렬 알고리즘은 속도는 빠른 대신 입력으로 사용되는 서열의 길이에 따라 메모리의 사용량이 크게 좌우되는 특징이 있다. 따라서 효율의 극적 향상을 위해서 원본 문서의 크기를 줄여야 할 필요가 생기게 되는데, 다음 절에서 본 논문에서 사용된 여러 가지 한글 문서 축약 방법을 소개하였다.

3.2 문서 축약법의 개발

기존의 문서 유사성 측정 방법의 가장 큰 문제점은 문서의 크기가 커졌을 경우 너무 비효율적이라는 것과 한글의 자유로운 어형 변화를 적절히 고려하지 못한다는 것이다. 문서의 크기는 커지면 커질수록 문서 전체를 비교하는 데에 많은 비용이

들게 되고, 이에 따라 수십, 수백개의 문서 상호간 모두 비교를 하는 데에 엄청난 시간을 소요하게 된다. 또한 한글의 자유로운 어형 변화는 단어별로 표절 여부를 검사할 때 장애 요인이 된다. 이와 같은 이유로 원본 한글 문서의 특징은 살리면서 어형의 단순한 변화같은 표절 기법에 잘 견딜 수 있고, 데이터의 크기는 많이 줄일 수 있는 축약 방법이 필요하게 되었다. 본 논문에서는 데이터의 크기를 확연히 줄이고, 본문보다도 표절 기법들에 영향을 적게 받을 것이라 예상되는 여러 가지 방법을 테스트해 보았다. 크게 다섯가지 방법을 사용하여 수집한 데이터 집합에 적용하여 실험하였는데, 각각은 다음과 같다.

1.축약타입A - 원본 문서에서 공백과 기호를 제거하여 순수한 한글만으로 축약하였다.
2.축약타입B - 각 단어의 형태소를 분석한 뒤 동사와 명사만 추출하여 축약하였다.
3.축약타입C - 동사의 첫 2글자와 명사만 추출하였다.
5.축약타입D - 동사, 명사, 형용사, 부사의 첫글자를 추출하였다. 어절의 첫단어만 추출한 것과 거의 일치한다.
4.축약타입E - 동사와 명사의 첫글자만을 추출하였다. 가장 높은 축약율을 가진다.

본 논문에서는 축약을 위해 어절의 형태소를 분석하는 데 (주)나라인포테크[3]의 <형태소 분석 시스템> 을 프로그램을 사용하였다.

4. 표절 탐색 수행 및 평가

4.1 테스트 데이터 준비

본 논문에서는 앞 절에서 언급한 5가지 한글 축약방법들을 이용하여 원본 문서를 축약한 뒤 지역정렬로 표절검사를 수행하고자 하는 수십개의 문서간의 상호 유사도를 측정할 수 있는 시스템을 개발하였다. 그리고 이 시스템의 테스트를 위해 기본적이고 수동적인 방법으로 표절 여부의 테스트를 이미 모두 끝낸 다음과 같은 여러 개의 데이터 집합에 대하여 시스템의 성능을 측정해 보았다.

1. 파리의 연인 관련 기사[1] - 집합-(가)
2. 가시고기 독후감[3] - 집합-(나)
3. 갈매기의 꿈 독후감[3] - 집합-(다)
4. 보아활동 수기[3] - 집합-(라)
5. 올드보이 감상평[3] - 집합-(마)
6. 어린왕자 독후감[3] - 집합-(바)

각 데이터 집합의 구성은 다음과 같다.

	문서 개수	비교 쌍	표절 쌍	표절 비율	크기(KB)
가	24	276	33	12.0%	53
나	25	300	12	4.0%	88
다	25	300	9	3.0%	157
라	130	8385	33	0.4%	570
마	16	120	7	5.8%	60
바	24	276	16	5.8%	135

표 1. 각 데이터 집합의 개수, 표절 정보 및 크기

4.2 테스트 데이터의 축약

앞에서도 언급하였듯이 본 연구에서 제시하는 다수의 한글문서간 표절 검사에 있어서 가장 핵심적인 요소는 본문을 어떻게 그리고 얼마나 축약을 하는가이다. 앞에서 언급한 다섯가지 축약방법으로 가~바의 여섯 개의 데이터 집합을 축약하였는데, 그 결과를 표 2와 표 3에 나타내었다.

데이터 크기가 각각의 방법에 따라 20% ~ 80% 비율로 줄어드는 것을 확인할 수 있다. 지역 정렬의 time complexity가 $O(N^2)$임을 상기하여 볼 때 프로그램의 수행시간은 축약 정도의 제곱으로 더 좋아지게 됨을 쉽게 예측할 수 있다. 또한 각 단어의 핵심 형태소를 추출함으로써 "하였다"를 "했다"등과 같이 기계적인 작업으로 변형 가능한 표절기법에 강건(robust)하게 대응할 수 있음을 알 수 있다.

	축약A	축약B	축약C	축약D	축약E
가(53 KB)	42(KB)	26	23	13	11
	81%	50%	44%	25%	21%
나(88 KB)	73	41	36	24	18
	83%	47%	41%	27%	20%
다(157 KB)	129	75	65	45	33
	82%	48%	41%	29%	21%
라(570 KB)	481	281	248	164	123
	84%	49%	44%	29%	22%
마(60 KB)	48	27	24	16	11
	80%	45%	40%	27%	18%
바(135 KB)	110	62	53	39	28
	81%	46%	39%	29%	21%

표 2 . 각 데이터 집합의 축약된 크기 및 축약 비율

	축약A	축약B	축약C	축약D	축약E
평균 축약 비율	82.0%	47.4%	41.6%	20.5%	27.5%

표 3 . 각 방법의 평균적 축약 비율

4.3 축약된 데이터의 효율성

데이터가 아무리 좋은 효율로 축약되었다고 하더라도 본래 문서의 특성을 잃어버리게 되어 표절여부를 제대로 검출해 내지 못한다면 아무 소용이 없는 일이다. 각 축약 방식의 원본 문서의 특징에 대한 무결성을 테스트하기 위해 각 축약된 문서들간의 표절 여부와 실제 원본 문서의 표절 여부를 테스트해 보았다. 각 문서집합을 각 축약방법으로 축약한 뒤 축약된 문서에 대해 지역 정렬 값을 구하여 그 값에 따른 sensitivity(SN)와 specificity(SP)를 그래프로 그려보았다. SN과 SP는 식 1과 같이 정의되는 값이다.

SN = 판별해낸 실제 표절 개수 / 실제 표절 수
SP = 판별해낸 실제 표절 개수 / 표절로 판별한 수 - 식 1

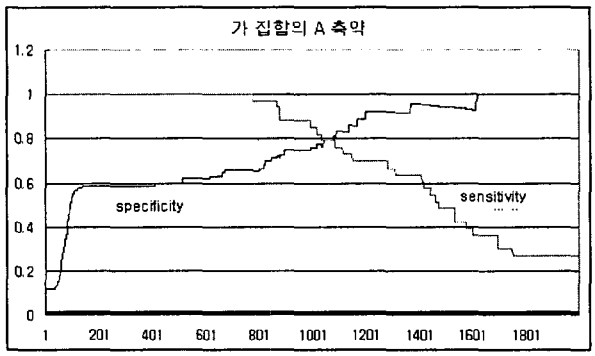

그림 1 . A 축약한 집합-(가)의 지역정렬 값에 따른 SN, SP

A 축약 방법은 원본 문서에서 공백과 기호만 없앤 것으로, 사실상 원본과 같다고 할 수 있고 따라서 그림 1은 가 원본 문서의 특징을 나타낸다고 할 수 있다. 한편 집합-(가)를 25% 정도의 가장 높은 축약율을 가지는 D 방법과 E 방법으로 축약한 후에 지역정렬하였을 경우의 sensitivity와 specificity는 그림 2, 3과 같은 형태를 가진다. 그림 2, 3으로 분석할 수 있는 핵심적인 내용은 이 그림들이 원본 문서의 특성을 나타내는 그림 1을 x축으로 축소한 형태의 그림이라는 것이다. 즉 한글문

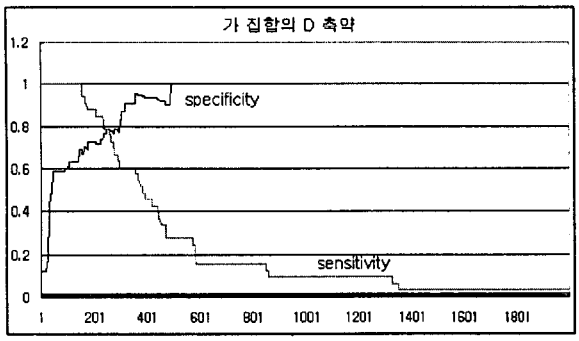

그림 2 . D 축약한 집합-(가)의 지역정렬 값에 따른 SN, SP

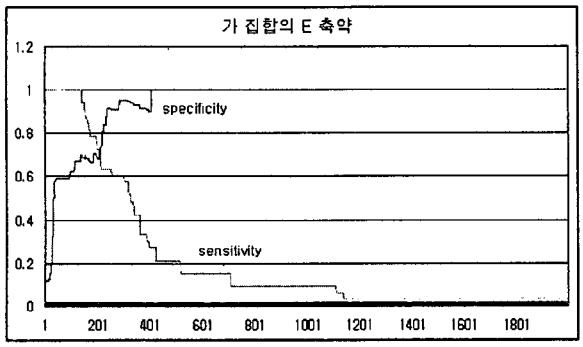

그림 3 . E 축약한 집합-(가)의 지역정렬 값에 따른 SN, SP

서의 핵심 내용인 형태소를 기준으로 축약을 한다면 25% 정도로 축약을 하더라도 문서의 유사성에 관한 특징이 변하지 않는다는 것을 의미한다. 그리고 sensitivity와 specificity의 교점을 기준으로 계산해 보면 그래프의 축약 정도는 데이터의 축약 정도와 매우 유사함을 확인할 수 있다. 따라서 원본 데이터를 동사와 명사의 형태소 일부만 추출하여 표절 여부를 검사하는 것은 데이터 유실에 관계없이 거의 오류가 없는 방법이라고 보는 것은 옳은 판단이다.

테스트 결과 집합-(가) 뿐만 아니라 나머지 모든 집합들이 축약율에 따라 그래프가 비슷한 형태로 x축으로 축소됨을 확인할 수 있었다. 그림 4, 5는 집합-(라)의 A축약과 22%의 축약율을 가지는 E축약에 대한 그래프들이다.

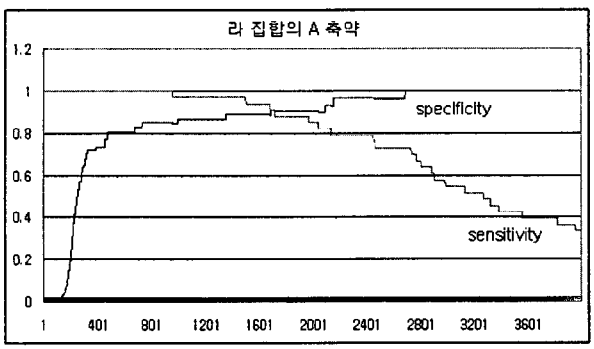

그림 4 . A 축약한 집합-(라)의 지역정렬 값에 따른 SN, SP

집합-(가)와 마찬가지로 가장 높은 문서의 축약율에 따라 표절 여부를 판별하는데 사용되는 원본 문서들의 유사성에 관한 성질이 그대로 축소됨을 확인할 수 있다.

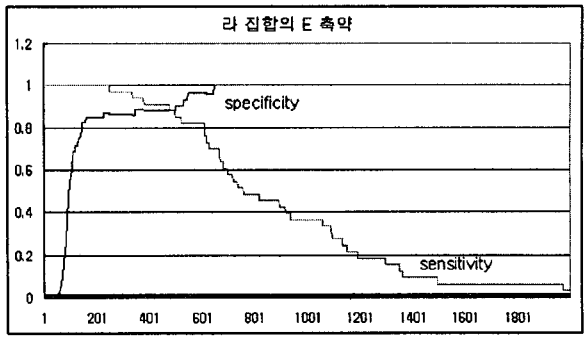

그림 5 . E 축약한 집합-(라)의 지역정렬 값에 따른 N, SP

그래프 분석 결과 SN과 SP 값이 1이 되는 위치나 1에서 벗어나는 위치, SN과 SP의 교점 등과 같은 주요 포인트는 문서 집합의 종류에는 비교적 적은 영향을 받고, 축약 방법에 크게 영향을 받는다는 것을 볼 수 있었다. 그림 1, 그림 4에서 보듯 문서가 A축약되었을 경우의 SN은 800~900 정도에서 감소하기 시작하며 SP는 2000 정도에서 1에 근접함을 확인할 수 있다. 또한 그림 2, 그림 3, 그림 5에서 보듯 D, E 축약의 경우에는 200의 지점에서 SN이 감소하기 시작하며, 600 지점에서 SP가 1에 근접함을 알 수 있다. 이것은 지역정렬이 찾아낸 표절 구간이 나타내는 특성으로, 표절구간의 길이는 문서의 종류에 영향을 받기보다는 각 축약정도에 따라 달라질 것이라는 것을 감안해보면 당연한 결과이다. 그 결과로 여기서 테스트한 일반 과제물 문서에서 D, E 축약의 경우 지역정렬 값이 600이 넘으면 표절로 봐도 무방하며, 200~600정도의 값을 가지는 것은 육안으로 판별할 필요성이 있음을 알 수 있다.

5. 결론 및 향후과제

본 논문에서는 한글의 특성을 고려한 효과적인 문서 표절검사 방법에 대해 살펴보았다. 여러 가지 실험을 거쳐 내린 결론은 다음과 같다.

- 한글의 특성상 형태소 분석을 통한 주요 어간의 분석이 시간과 공간면에서 효율적이다. 특히 전체 크기를 1/4로 줄여도 민감도나 정확도면에서 성능을 유지한다.
- 보통의 한글 과제물 문서인 경우 본 논문에서 제시한 D, E 축약의 방법으로 표절 여부를 검사한 경우 지역정렬 값이 600 이상인 문서는 표절로 판정해도 무방하며 200 이상 600 이하는 사용자의 육안검사를 거치는 것이 좋다.

표절로 확정된 문서 외에 표절 문서로 강하게 의심되는 문서에 대해서는 사람의 최종 확인을 거쳐야 할 필요가 있다. 본 논문에서는 문서간의 유사도만을 구하는 시스템을 구축하였는데, 향후에는 표절로 의심되는 문서들을 상호 비교, 가시화해주는 툴을 구현하여 사람이 보다 손쉽게 표절 의심문서를 확인할 수 있는 비주얼한 시스템을 구현할 계획에 있다.

참고자료(References)

[1] 전명재, 이평준, 조환규, "분할정령 방식을 이용한 지역정렬과 이를 이용한 소스코드 표절 탐색 기법", 한국정보과학회, 2004.4.
[2] Michael j. Wise, "YAP3: Improved Detection of Similarities in Computer program and other texts"
[2] 형태소 분석기, (주)나라인포테크
[3] 부산과학고등학교 과제물 협조, "http://www.bsa.hs.kr"
[4] WCopyFind, "http://www.plagiarism.phys.virginia.edu/"
[5] Eve2, "http://www.canexus.com/eve/"