CopyCheck: 한글문서 표절검사 소프트웨어

박소영○, 장은서, 권도형, 강승식
국민대학교 컴퓨터공학부
park-soyeong@nate.com, akdangz@kookmin.ac.kr, kdhlook@naver.com, sskang@kookmin.ac.kr

CopyCheck: Korean Document Plagiarism Detection System

So-Yeong Park○, Eun-seo Jang, Do-Hyung Kwon, Seung-Shik Kang
School of Computer Science, Kookmin University

요 약

본 논문에서는 대학의 과제물이나 학위 논문 또는 회사의 입사지원서, 자기소개서와 같은 문서에 대하여 표절검사에 활용할 수 있는 소프트웨어인 CopyCheck를 설계 및 개발하였다. CopyCheck는 표절검사 방법을 빠른 검사와 정밀 검사를 두어 보다 사용자가 편리하게 사용할 수 있도록 하였다. 표절검사를 진행한 후, 전체보기와 구간보기, 표절구간 시각화의 3가지 방법을 통해 사용자가 다양한 방법으로 표절 문서를 파악할 수 있도록 도와준다. 또한, 표절검사 결과를 저장할 수 있도록 하여 사용자가 언제든지 다시 볼 수 있도록 하였다.

주제어: 표절, 한글 처리, 표절검사, 한글문서 표절검사 소프트웨어

1. 서론

21세기 정보화시대가 도래하면서 컴퓨터의 발달이 급속도로 진행되고 있다. 이로 인한 매체의 활성화로 누구나 손쉽게 정보를 공유하고 가져올 수 있게 되어 표절이 보다 빈번하게 이루어지고 있다. 특히 학교 안에서의 과제, 논문 등의 표절이 빈번하게 일어나고 있으며, 더 나아가 회사의 입사지원서 또는 자기소개서와 같은 문서에 대해서도 표절이 빈번하게 일어나고 있다. 이로 인해 표절은 사회 문제로 더욱더 크게 대두되는 추세가 지속되고 있다. 이러한 결과로 문서간의 표절 검사에 대한 필요성과 수요 또한 증가하고 있다. 그러나 수작업으로 문서간의 표절 검사를 하는 것은 인적·시간적 비용이 너무 크다는 단점이 존재한다.

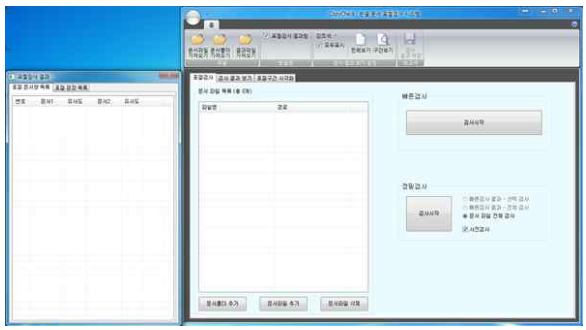
이러한 추세에 따라 한글문서 표절검사 방법에 대한 연구가 활발하게 진행되고 있으며 이미 공개되어 있는 한국어 표절 검사 소프트웨어들도 다수 존재한다. 그러나 대부분의 한국어 표절 검사 소프트웨어들은 수행시간이 지나치게 오래 걸리거나 표절의 가능성이 낮은 문장에 대해서도 발견해낸다는 문제가 있다.

본 논문은 표절 검사 대상 문서 군에서 표절이 의심되는 문서 군을 군집화 하고, 표절 의심 문서들 간의 문자열 일치구간을 빠르고 정확하게 찾아낼 수 있는 표절 검사 소프트웨어인 CopyCheck를 설계 및 개발하였다.[1] CopyCheck는 표절 검사 방법을 빠른 검사와 정밀 검사를 두어 보다 사용자가 편리하게 사용할 수 있도록 하였다. 표절검사를 진행한 후, 전체보기와 구간보기, 표절구간 시각화의 3가지 방법을 통해 사용자가 다양한 방법으로 표절 문서를 파악할 수 있도록 도와준다. 또한, 표절검사 결과를 저장할 수 있도록 하여 사용자가 언제든지 다시 볼 수 있도록 하였다.

2. CopyCheck

우리는 서론에서 소개한 이미 공개되어 있는 소프트웨어들의 문제점을 극복하고자 새로운 표절검사 소프트웨어인 CopyCheck를 설계 및 개발하였다. 본 절에서는 CopyCheck version 1.0의 구조와 기능들을 상세히 설명한다.

2.1 CopyCheck의 구조



[그림 1] CopyCheck 소프트웨어 실행 화면

CopyCheck의 초기화면은 [그림 1]과 같이 main화면과 main화면 좌측에 위치한 '표절검사 결과' 창으로 구성된다. main화면의 상단은 홈 탭으로 구성되어 있으며, 홈 탭의 하단은 표절검사, 검사 결과 보기, 표절구간 시각화로 총 3개의 탭으로 구성되어 있다. 초기 소프트웨어를 실행하면 표절검사 탭에서 시작하게 된다.

[그림 2] 홈 탭 화면

홈 탭의 메뉴는 [그림 2]와 같이 파일, 창설정, 검사

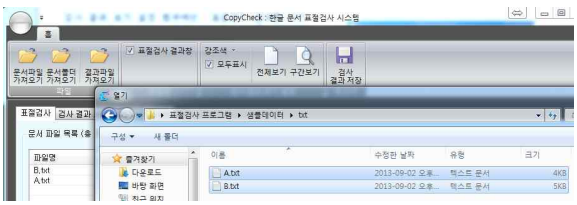
결과 보기 설정, 보고서 4개의 범주가 있다. 파일 범주에는 표절검사를 위한 문서파일 가져오기, 문서폴더 가져오기, 결과파일 가져오기를 선택하기 위한 버튼들이 있다. 창설정 범주에는 main화면의 좌측에 위치한 표절검사 결과창 활성화를 설정할 수 있는 체크박스가 있다. 검사 결과 보기 설정 범주에는 표절검사 결과를 보기 위한 강조색, 모두표시, 전체보기, 구간보기를 선택하기 위한 버튼들이 있다. 보고서 범주에는 표절검사 수행 후, 결과 파일을 저장하는 검사 결과 저장 버튼이 있다.

2.2 표절검사 문서 파일 선택

표절 검사할 문서 파일을 선택하는 방법은 파일 단위로 선택하는 방법과 폴더 단위로 선택하는 방법 2가지가 있다. 또한 파일을 선택하는 버튼이 상단과 하단에 2가지가 있으며, 상단에 위치한 버튼은 기존의 파일 목록을 없애고 새롭게 목록을 생성하는 기능이며, 하단에 위치한 버튼은 기존의 파일 목록에 추가되는 기능이다.

2.2.1 파일 단위 선택

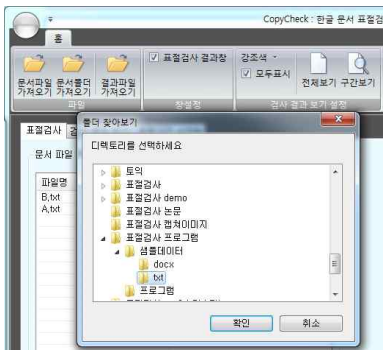
표절 검사할 파일을 파일 단위로 선택하는 방법은 상단의 문서파일 가져오기 또는 하단의 문서파일 추가 버튼을 클릭하는 것이다. 클릭 시 파일을 선택할 수 있도록 탐색기 창이 나타나며, 표절 검사할 문서 파일들을 선택하고 열기 버튼을 누르면 된다. [그림 3]은 샘플 데이터 폴더 아래의 txt폴더에서 A.txt와 B.txt를 선택한 예이다.



[그림 3] 파일 단위 선택 화면

2.2.2 폴더 단위 선택

표절 검사할 파일을 폴더 단위로 선택하는 방법은 상단의 문서폴더 가져오기 또는 하단의 문서폴더 추가 버튼을 클릭하는 것이다. 클릭 시 폴더를 선택할 수 있도록 탐색기 창이 나타나며, 표절 검사할 문서 폴더를 선택하고 확인 버튼을 누르면 된다. [그림 4]는 샘플 데이터 폴더 아래의 txt폴더를 선택한 예이다.



[그림 4] 폴더 단위 선택 화면

2.3 표절검사 모드

표절검사 모드는 빠른 검사와 정밀 검사가 있다. 빠른 검사는 검사 속도는 빠르지만 정밀 검사에 비해 표절의 심구간을 파악하는 정확도가 떨어질 수 있다. 반면 정밀 검사의 경우는 검사 속도가 느린 편이지만 표절의심구간을 파악하는 정확도가 빠른 검사에 비해 높다.

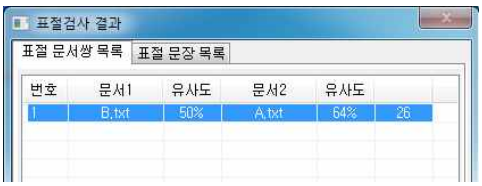
빠른 검사는 검사 후 '빠른 검사 결과-선택 검사'와 '빠른 검사 결과-전체 검사' 2가지 모드가 있다. 먼저 '빠른 검사 결과-선택 검사'는 빠른 검사를 통해 표절이 의심되는 문서 파일들을 추려내고, 사용자가 원하는 문서 파일들을 선택하여 정밀 검사하는 기능이다. 다음으로 '빠른 검사 결과-전체 검사'는 빠른 검사를 통해 추려낸 문서 파일 전체를 정밀 검사하는 기능이다. 문서 파일 전체 검사의 버튼을 체크한 후, 정밀 검사의 검사 시작 버튼을 누르면 사용자가 업로드한 문서 파일 전체에 대해 정밀검사를 진행한다.

[그림 5] 표절검사 모드

2.3 표절검사 결과보기

표절검사를 진행한 후, 표절 의심 문서 파일들이 존재하면 [그림 6]과 같이 표절검사 결과창의 표절 문서쌍 목록 탭에 한 행마다 한 쌍의 의심되는 파일들의 이름과 유사도, 의심되는 문장 수가 출력된다. 상세하게 보고 싶은 한 행을 더블 클릭하면 자동으로 표절검사 결과창은 [그림 7]과 같이 표절 문서쌍 목록 탭에서 표절 문장 목록 탭으로 이동하게 되고, main화면은 표절검사 탭에서 검사 결과 보기 탭으로 이동한다.

표절검사 결과는 사용자가 업로드한 문서 파일들 중에서 표절 의심 문서 파일들에 대해 1:1 비교 방식으로 나타내며, 결과를 보는 방법은 검사 결과 보기와 표절구간 시각화 2가지가 있다. 검사 결과 보기는 표절 의심 문서 파일들 중 각각의 의심되는 문장을 보여준다. 또한 표절구간 시각화는 검사 결과 보기와는 달리 표절 의심 문서 파일들의 내용이 보이지 않고, 표절 의심이 되는 구간의 분포도를 보여준다. 자세한 내용은 각각의 범주에서 설명하도록 한다.



[그림 6] 표절검사 결과 - 표절 문서쌍 목록

[그림 7] 표절검사 결과 - 표절 문장 목록

2.3.1 검사 결과 보기

검사 결과 보기는 전체보기와 구간보기 2가지 방법이 있다. 이는 홈 탭 메뉴에 위치한 검사 결과 보기 설정 범주에서 선택할 수 있다. 전체보기와 구간보기는 [그림 7]과 같은 표절검사 결과창에서 각 행을 클릭하면 해당 행의 표절의심구간에 대해 비교하여 볼 수 있다. 전체보기는 [그림 8]과 같이 표절 의심 문서 파일의 내용 전체를 보여주며, 그 중 의심되는 문장을 강조색을 통해 보여준다. 반면 구간보기는 [그림 9]와 같이 표절 의심 문서 파일 중 의심되는 문장 별로 비교하여 보여준다.

[그림 8] 검사 결과 보기 - 전체보기

[그림 9] 검사 결과 보기 - 구간보기

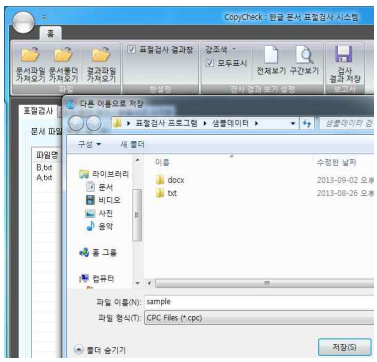
2.3.2 표절구간 시각화

표절구간 시각화는 앞서 설명한 검사 결과 보기와는 달리 실제 의심되는 문서 파일의 내용이 보이지 않는다. 다만, [그림 10]과 같이 표절 의심 문서 파일에 대해서 표절의심구간인 곳들을 표시하여 사용자가 표절의심구간의 분포를 한 눈에 확인할 수 있다는 이점이 있다.

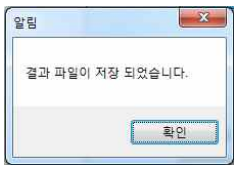
[그림 10] 표절구간 시각화

2.4 표절검사 결과 저장

표절구간 결과를 저장하고 싶다면 보고서 범주의 검사 결과 저장 버튼을 클릭한다. 이때 [그림 11]과 같이 탐색기 창이 뜨면 원하는 경로를 선택한 다음 원하는 파일 이름을 입력하고 저장 버튼을 누른다. 이와 같이 진행하면 [그림 12]와 같이 알림창이 뜬다.



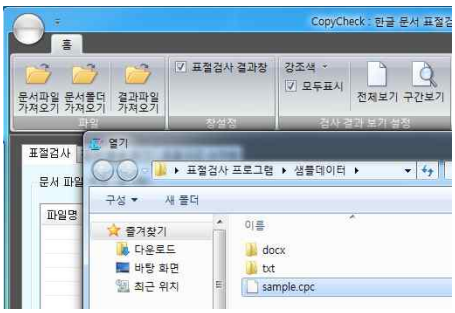
[그림 11] 표절검사 결과 저장



[그림 12] 결과 저장 알림창

2.5 표절검사 결과 불러오기

이전에 표절검사를 진행한 결과를 다시 보고 싶다면, [그림 13]과 같이 파일 범주의 결과파일 가져오기 버튼을 클릭한다. 이때 주의할 점은 이전에 검사를 진행하고 나서 검사 결과 파일(*.cpc)을 저장하였을 때 가능하다.



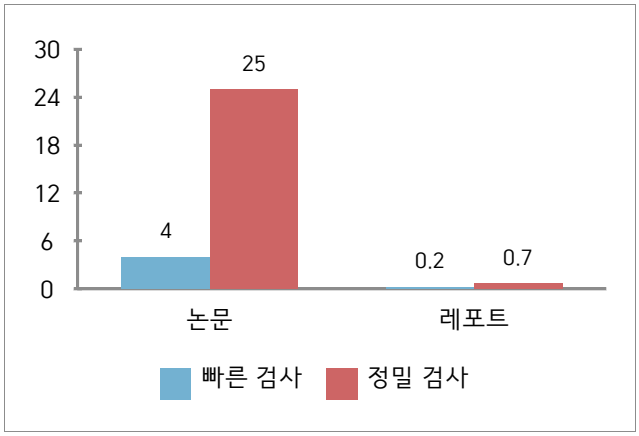
[그림 13] 표절검사 결과 파일 가져오기 화면

3. 실험 및 평가

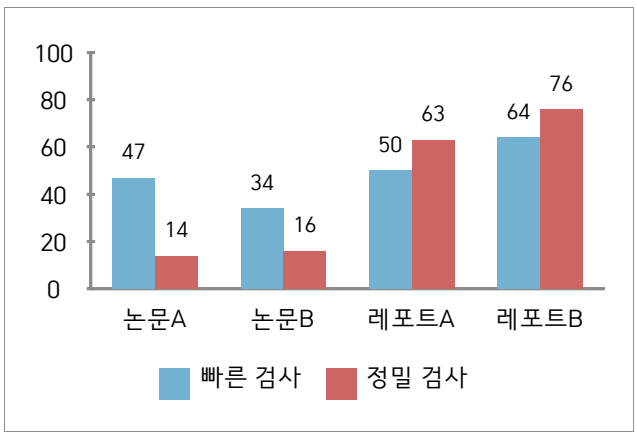
본 논문에서 설계 및 개발한 한글문서 표절검사 소프트웨어를 이용하여 실험을 진행한다. 실험은 문서를 1:1로 비교하여 진행하며 빠른 검사와 정밀 검사에 대해 각

각 표절검사 수행시간과 표절의심 문서간의 유사도를 측정한다. 실험에서 사용되는 샘플 문서의 종류는 총 2가지로 논문과 학생들의 레포트로 구성하였다.

[그림 14]는 2가지의 샘플 문서의 종류에 대해서 각각의 표절검사 모드인 빠른 검사와 정밀 검사의 수행시간을 측정하여 비교한 것이며 [그림 15]는 표절의심 문서간의 유사도를 비교한 것이다.



[그림 14] CopyCheck: 한글문서 표절검사 소프트웨어 수행시간 측정



[그림 15] CopyCheck: 한글문서 표절검사 소프트웨어 표절의심 문서간의 유사도 측정

4. 결론

본 논문에서는 한글문서 표절검사 소프트웨어인 CopyCheck를 설계 및 개발하였다. CopyCheck는 사용자의 편의성을 높이기 위해 다양한 표절검사 방법과 표절검사 결과 보기 방법을 제공한다. 또한, 표절검사 결과를 저장 또는 불러오기 기능을 두어 사용자가 언제든지 표절검사 결과를 다시 볼 수 있도록 한다.

실제 CopyCheck를 사용·도입하여 표절 의심문서를 판별해냄으로 요즘 사회문제로 대두되고 있는 콘텐츠의 지적재산권을 보호할 수 있다. 또한, 수작업으로 표절검사를 진행할 때에 비해 쉽고 빠르게, 정확하게 표절 의심문서를 판별할 수 있다는 장점이 있다. 궁극적으로는 문서간의 표절을 줄일 수 있을 것이라고 기대한다.

참고문헌

[1] 장은서, 권도형, 김낙원, 박소영, 강승식, “CopyCheck : 한국어 표절 검사 소프트웨어”, 제24회 한글 및 한국어 정보처리 학술대회, pp.117-118, 2012.

[2] 류창건, 김형준, 조환규, “한글 말뭉치를 이용한 한글 표절 탐색 모델 개발”, 정보과학회논문지: 컴퓨팅의 실제 및 레터 제14권 제2호, pp.231-235, 2008.

[3] 지혜성, 조준희, 임희석, “한국어 문장 표절 유형을 고려한 유사 문장 판별.”, 컴퓨터교육학회논문지 제13권 제6호, pp.79-89, 2010.

[4] 안병렬, 김문현, “문서를 위한 표절 탐지 소프트웨어에 관한 연구.”, 한국퍼지 및 지능소프트웨어학회 2006 년도 춘계학술대회 학술발표논문집 제16권 제1호, pp.413-415, 2006.

[5] 박선영, 조환규, “성분 정렬을 이용한 한글 유사문서 탐색 방법.”, 한국정보과학회 2011 한국컴퓨터종합학술대회 논문집 제38권 제1호, 2011.