한글 말뭉치를 이용한 한글 표절 탐색 모델 개발

류창건°, 김형준, 박병준, 최혜정, 조환규
부산대학교 컴퓨터공학과
{ckryu,hjkim83}@pearl.cs.pusan.ac.kr, {phantomman,prettycomic}@naver.com, hgcho@pusan.ac.kr

Construction of Text Plagiarism Detection Model using Korean Corpus Data

Chang-Keon Ryu°, Hyong-Jun Kim, Byong-Jun Park, Hae-Jeong Choi, Hwan-Gue Cho
Dept. of Computer Engineer, Pusan National University

최근 들어 각종 창작물의 표절사건이 갈수록 사회문제화 되고 있다. 이미 영어권에서는 일반 영어문서의 표절에 대한 연구[1,2,3,4,5]가 오래전부터 이루어져 왔지만 한글은 그 구조적 특성으로 인하여 아직 많은 연구가 이루어져 있지 못하다. 영어권에서 사용되는 표절탐색 알고리즘은 조사의 다양한 변화를 허용하는 한글문서에서 큰 효과를 보이지 못하고 있다. 따라서 한글의 특성에 맞는 한국어 전용 표절 탐색 시스템의 개발이 매우 시급하게 요청되고 있다. 그런데 아직 한글 표절 탐지 시스템은 널리 알려진 것이 없을 뿐만 아니라 표절 시스템 개발에 필요한 표준화 된 실험 데이터도 없는 상황이다. 이에 본 논문은 한글 문서에 맞는 표절 탐색 시스템을 제안하고, 한글 정보화 사업을 위해 제작한 한글 말뭉치를 이용하여 보다 새로운 한글문서 표절탐색 방법론을 제안하고자 한다.

본 논문에서 제시하는 한글 표절 탐색 시스템은 'DEVAC(Document EVolution Analyzing Center)' 시스템[6]이라 불리며, 한글의 특성에 맞게 시스템이 동작할 수 있도록 'fingerprint' 방식[7]과 'BLAST' 방식[8]을 혼합하여 사용하고 있다. 'fingerprint' 방식은 인덱스 검색 방식으로 빠른 탐색을 할 수 있는 장점을 가지고 있고,'BLAST' 방식은 대용량 데이터에서 유사한 부분을 빠르게 찾을 수 있는 장점을 가지고 있어 DEVAC 시스템은 대용량 한글 문서에서도 유연하고 빠른 탐색이 가능한 특징을 가지고 있다. DEVAC 시스템[10,11]은 문서의 전처리 단계, 표절을 탐색하는 단계, 표절 판정 및 결과 리포트 단계로 나누어서 수행된다. 전처리 단계에서는 여러 가지 입력 문서들을 DEVAC 시스템에서 정의한 표준 포맷 형식으로 변환하는 작업과 문서에 인덱스를 두어 사전 구조로 만드는 작업을 수행한다. 표절을 탐색하는 단계는 사전 구조를 비교하여 공통 앵커를 찾는 작업과 불용어 제거 작업을 거쳐 실제 표절 탐색을 하는 작업을 수행한다. 마지막으로 표절 판정 및 결과 리포트 단계는 한글 말뭉치를 이용한 실험을 통해 구해진 확률 모델과의 비교를 통한 표절 판정 작업과 결과 리포트를 작성하는 것으로 마무리 된다.

한글 표절 탐색 시스템의 개발과 성능 측정을 위해서는 표절 영역이 전혀 없는 정규화된 문서가 필요하다. 이에 본 논문은 한글 말뭉치[9]를 통하여 한글 문서간의 안정적인 유사도 측정을 할 수 있는 확률 모델을 제시하고 그에 근거한 새로운 표절 판정 방법을 제시한다. 한글 말뭉치를 입력 데이터로 하여 DEVAC 시스템을 수행한 결과, 유사도 그래프가 푸아송(poisson) 그래프와 유사하게 나타났다. 약간의 오차는 존재하나 이 부분은 추후에 파라미터 변화를 통하여 더욱 정밀하게 맞출 계획이다. 표절을 제거한 말뭉치 데이터를 통한 유사도 그래프 값이 푸아송 그래프를 따르므로 이 확률 모델을 이용하여 입력 문서들 사이의 표절 판정을 할 수 있다. 확률 모델에서 유사도 값이 100이 넘을 경우가 확률이 10%라

가정하자. 이때, 입력 문서의 표절 유사도 값이 100으로 나타나면 이 입력 문서들 사이에 표절이 이루어졌을 가능성이 90% 이상이 되며, DEVAC 시스템은 이 확률 값을 사용자에게 알려준다. 기존의 표절 탐색 시스템은 객관적인 기준을 제시하지 않고, 개발자가 자의적으로 기준을 정하거나 사용자의 판단에 맡기는 경우가 많지만 DEVAC 시스템은 표절에 관한 확률 모델을 제시하여 문서들 사이의 표절 여부를 객관적으로 제시한다.

본 시스템의 유수성과 실용성을 위하여 인터넷 검색으로 얻은 문서와 리포트를 대상으로 한 표절 탐색 실험을 수행하였다. 구글 뉴스에서 동일한 키워드를 통해 검색된 10개의 문서들을 대상으로 표절 탐색을 수행한 결과 4개의 뉴스들 사이에서 높은 표절 유사도 값이 나타났고, 또 다른 2개의 뉴스들 사이에서도 높은 표절 유사도 값이 나타났다. 뉴스의 출처를 살펴 본 결과, 3개의 뉴스가 연합뉴스를 기사의 출처를 밝히지 않고 표절한 사실을 알 수 있었고, 한 언론사가 회사 이름을 영문 표기와 한글 표기를 다르게 하여 다른 사이트에 기사를 올린 사실을 알 수 있었다. 구글 뉴스는 다른 포털 사이트와 달리 기사의 중복을 제거하는 기능을 갖추고 있었으나 이 중복 제거 기술로 제거되지 않은 중복 기사들이 많은 것을 알 수 있다. 만약 DEVAC 시스템을 포탈 뉴스의 필터링으로 사용하게 된다면 동일 기사를 재전송하는 언론사들의 검색 어뷰징 현상을 방지할 수 있고, 검색 결과의 신뢰성을 높일 수 있을 것이다. 또 다른 실험으로 한 학급에서 작성한 55편의 독후감을 표절 탐색한 결과, 9개의 문서 사이에 표절이 이뤄진 것을 발견할 수 있었다. 유사도가 높게 나타난 두 문서를 직접 비교해 본 결과, 중간 중간 어절의 삭제와 삽입 및 치환이 이루어졌으나 두 문서는 분명히 동일한 의미를 지닌 문장들을 가지고 있었다. 실험 결과를 살펴보면 실제 표절 문서를 매우 정확하게 찾아 낼 뿐 아니라 여러 분야에서 다양하게 사용될 수 있음을 알 수 있었다.

본 논문은 한글 문서의 표절을 탐색할 수 있는 표절 탐색 시스템을 제안하였고, 한글 말뭉치를 통한 실험을 통해 시스템의 성능을 증명하였다. 또한 한글 말뭉치를 이용한 실험을 통해 확률 모델을 만들었고, 이 모델을 기준으로 제시하여 표절 여부를 객관적으로 판별하도록 하였다. 또한, 여러 분야에서의 응용 예를 보여 표절 탐색 시스템의 실용성과 정확성을 보였다.

참고문헌

[1] Turnitin. http://www.turnitin.com/

[2] CloneChecker: A Software Plagiarism Detector. http://ropas.snu.ac.kr/n/clonechecker/

[3] Geoff Whale. Plague: Plagiarism detection using program structure. Department of Computer Science, University of New South Wales, May 1988.

[4] David Gitchell and Nicholas Tran. Sim: a utility for detecting similarity in computer programs. In SIGCSE '99: The proceedings of the thirtieth SIGCSE technical symposium on Computer science education, 266-270, 1999.

[5] Wise. YAP3: Improved detection of similarities in computer program and other texts. SIGCSEB: SIGCSE Bulletin, 28, 1996.

[6] Ryu Chang-Keon, Kim Hyong-Jun, Park Soo-Hyun, and Cho Hwan-Gue. DEVAC(Document EVolution Analyzing Center). http://devac.cs.pusan.ac.kr:8080/

[7] Schleimer, S., Wilkerson, D. S., and Aiken, A. Winnowing: local algorithms for document fingerprinting. In Proceedings of the 2003 ACM SIGMOD international Conference on Management of Data. 76-85. June 09 - 12, 2003

[8] Cameron, M., Williams, H. E., and Cannane, A. Improved Gapped Alignment in BLAST. IEEE/ACM Trans. Comput. Biol. Bioinformatics 1, 3, 116-129. Jul. 2004.

[9] 21세기 세종계획. http://www.sejong.or.kr/

[10] Ryu Chang-Keon. 한글 표절 탐색 시스템 개발을 위한 기초연구(1). Technical Report No. SYS07003DEVAC, Graphics application lab, Pusan University, Apr. 2007.

[11] Ryu Chang-Keon. 한글 표절 탐색 시스템 개발을 위한 기초연구(2). Technical Report No. SYS07004DEVAC, Graphics application lab, Pusan University, May. 2007.