표절 문서의 시각화를 위한 문서 진화 계통도 생성

김선영○ 박선영 조환규
부산대학교 정보컴퓨터공학부
{s.y.kim, parksy, hgcho}@pusan.ac.kr

Generating Phylogenetic Tree for Visualization of Plagiarized Documents

SeonYeong Kim○ Sun-Young Park, Hwan-Gue Cho
Computer Science & Engineering, Pusan National University

1. 서 론

문서 표절이 사회적 문제로 부각되면서 유사 문서 탐색 시스템에 대한 연구가 증가하였고 이러한 연구들에 대한 가시적인 성과가 나타나고 있다. 유사 문서 탐색 시스템은 어떤 문서 집합 내에서 표절이 이루어진 문서를 찾아내는 것을 목표로 한다. 하지만 문서 집합 내에서 표절이 이루어진 문서 쌍이 다수 존재할 경우 원 저작물과 표절물을 구분하여 찾는 것이 매우 어려우며, 표절의 흐름을 파악하는 것이 힘들다. 따라서 본 논문에서는 표절 탐색 시스템으로 찾은 탐색 결과를 바탕으로 하는, 표절 추적 흐름도를 생성하는 시스템을 제안한다. 제안한 시스템을 DeVAC[1]의 탐색 결과 모델에 적용하여, 사용자가 효과적으로 표절을 인식하고 그 흐름을 빠른 시간 내에 판단하는데 도움을 줄 수 있도록 한다.

2. 본 론

표절 탐색 시스템에서 사용하는 유사도 측정 방법은 상대 유사도 측정 방법과 절대 유사도 측정 방법으로 나눈다[2]. 대부분의 시스템이 두 가지 방법을 각각 혹은 복합적으로 사용하여 유사 문서 쌍을 탐색하며, 많은 경우 도표를 이용하여 유사한 순서대로 문서 쌍을 보인다. 이는 유사 정도를 파악하는데 도움이 되나 원 저작물과 표절물을 구분하기 힘들다는 문제가 있다. 본 논문에서는 생물학에서 사용하는 진화 계통도(Phylogenetic Tree)를 문서 집합에 적용하여, 각 문서를 유전체로 보고 그 진화의 흐름을 추적하는 방식으로 표절이 발생하는 정도와 표절의 방향을 추적하는 시스템을 제안한다. 이 시스템은 1:1 문서 표절 검사에 우수한 성능을 보이는 DeVAC에서 얻은 표절 결과를 활용하여 진화 계통도를 설계하였다. DeVAC은 두 문서의 유사성을 판단하는 기준으로 절대 유사도를 사용하며 보완적 측면 에서 상대 유사도도 함께 사용한다.

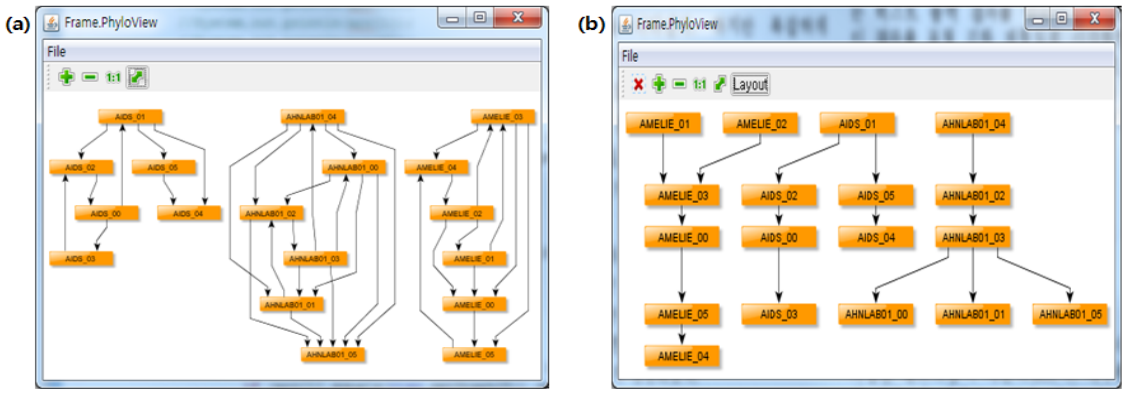

그림 1. 의도적으로 생성한 유사 문서에 대한 진화 계통도. 각 노드는 문서를 나타내며, 간선은 진화 관계를 나타낸다. (a)는 모든 유사 쌍에 대한 관계가 표현되어 있다. (b)는 (a)에서 사이클과 다중 진입점(multi entry)을 제거하기 위해 연결이 약한 간선을 삭제한 그래프.

진화 계통도를 생성하기 위하여 모든 문서를 노드로 만들고 모든 유사 관계에 대해 문서 간의 에지를 생성하여 그래프를 만든다. 의도적으로 생성한 표절 데이터를 앞서 언급한 방식으로 생성한 계통도로 나타내면, 그림 1 (a)와 같이 문서들이 그룹을 이루어 분할됨을 확인할 수 있는데, 이는 표절의 계통이 분명히 존재함을 의미한다. 이렇게 얻은 계통도는 관련된 모든 문서에 연결이 있으므로 한 눈에 흐름을 파악하기 어렵기 때문에 진화 계통도에서 사이클과 다중 진입점(multi entry)이 모두 제거될 때까지 유사성이 낮은 연결을 제거하는 과정이 필요하다. 이 작업을 수행하면 유사도가 높은 문서들 간의 연결만 남기 때문에 그림 1 (b)와 같이 정확한 표절 흐름의 파악이 가능하다. 또한 상위 연결이 더 유사하다는 것을 의미하고, 절대 유사도 개념을 활용하면 최상위 노드가 원저자일 가능성이 높고 하위 노드로 갈수록 표절자일 확률이 높다고 할 수 있다.

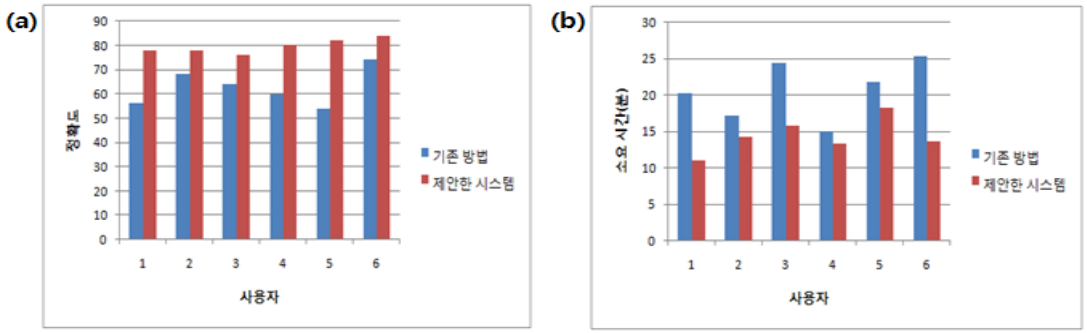

그림 2. 원 저작물/표절물 판정 및 표절 정도의 인식 결과 실험. 사용자는 각각 기존 방법과 제안한 시스템을 사용하여 원저자와 표절자를 구분하였다. (a)는 정답률을 비교한 결과이며, (b)는 판정하기 까지 소요한 시간을 비교한 결과이다.

원저자/표절자 판정 및 표절 정도의 인식에 표절 진화 계통도가 도움이 되는 정도를 평가하기 위해 6명의 피 실험자를 대상으로 500개 이상의 문서로 구성한 서로 다른 두 개의 문서 집합A, B에 대한 진화 계통도를 생성하여 표절 흐름을 판단하는 실험을 실시하였다. 실험 결과는 그림 2와 같다. 표절 계통도를 제공할 경우 기존에 비해 피 실험자가 표절 관계를 판단하는데 필요한 평균 시간은20.7분에서 14.3분으로, 표절 판정의 정확도는 62.67%에서 79.67%로 개선되었다. 주목할 만한 사실은, 제안한 시스템을 사용했을 때 대부분의 사용자가 진화 계통도가 보여주는 그대로 표절 여부를 판정했다는 사실이다. 이는 본 시스템의 원저자/표절자 구분의 정확도를 높일 경우, 매우 빠른 시간 내에 높은 정확도로 표절 관계 및 표절 흐름을 인식할 수 있음을 의미한다.

3. 결 론

대부분의 유사 문서 탐색 시스템은 찾아낸 유사 문서 표시 방법을 텍스트 기반의 결과로만 제공하여 사용자가 원 저작물과 표절물을 판별하기 매우 어렵다. 본 논문에서는 진화 계통도를 이용해 표절 검사 결과를 시각화하여 유사 문서의 흐름을 추적하는 시스템을 개발하였다. 실험 결과, 표절 문서의 원본을 찾는데 소요한 시간은 평균 6.4분, 판정의 정확
도는 평균 17% 개선되어, 본 시스템이 사용자가 표절 여부를 판단하는데 실질적인 도움을 줄 수 있음을 확인하였다.

참고 문헌

[1] 류창건, 김형준, 박수현, 박선영, 조환규, DeVAC(Document eVolution Analysis Center), http://devac.cs.pusan.ac.kr, 2010.

[2] Chang-Keon Ryu, Hyong-Jun Kim and Hwan-Gue Cho, A Detecting and Tracing Algorithm for Unauthorized Internet-News Plagiarism Using Spatio-Temporal Document Evolution Model, In Proc. of ACM SAC, pages 863-868, 2009.