문서를 위한 표절 탐지 시스템에 관한 연구

A Study on Plagiarism Detection System for Documents

안병렬[1], 김문현[2]

[1] 경기도 수원시 장안구 천천동 성균관대학교 컴퓨터공학과
E-mail: anbr0305@skku.edu
[2] 경기도 수원시 장안구 천천동 성균관대학교 컴퓨터공학과
E-mail: mhkim@ece.skku.ac.kr

요 약

디지털 시대에는 누구나 쉽게 정보에 접근 할 수가 있어 아주 간단하게 다른 사람의 정보를 불법 복제해서 무단으로 사용하는 경우가 증가하게 되었다. 이는 많은 투자와 노력으로 지식을 생성하는 일도 중요하지만 이를 관리하고 보호하는 일이 중요한 과제로 부상하고 있다는 것을 의미한다. 본 논문에서는 다른 사람의 지적 재산권을 침해하고 표절을 하여 사용했을 경우 이를 효과적으로 탐지하는 새로운 방법과 이론을 제시하고자 한다.

Key Words : Plagiarism, Detection System, Digital Contents Protection

1. 서 론

컴퓨터 기술의 향상과 정보의 중요성이 더해지면서 갈수록 지적재산권에 대한 침해와 표절이 증가하고 있다. 표절과 불법 복제가 성행하고 있지만 이에 대한 대처 방법과 연구가 국내외적으로 아직까지 미흡한 실정이다. 표절의 판별과 감정에는 일일이 사람들의 손을 거쳐야 하며 많은 시간과 자원의 소요가 뒤 따른다. 따라서 좀더 효율적인 방법론과 객관적이고 시스템적인 접근이 필요하다고 본다. 앞으로 이러한 분야에 많은 연구와 개발이 이루어 질 것으로 예상되며 이로 인해 많은 표절이나 불법 복제의 행위가 쉽게 검출될 것이고, 지적재산권의 분쟁 해결과 함께 피해를 최대한 줄일 수 있을 것이다.

수많은 문서 중에서 복제 여부를 판단하는 데는 많은 시간과 인력이 투입되고, 문서 하나 하나 마다 비교하여 유사율까지 결과가 산출되기 까지는 많은 복잡성이 존재하게 된다. 이에 주안점을 두고 어떻게 하면 많은 문서를 빠르게 비교하면서 신뢰성을 높일 수 있는 문서 검출 방법을 강구 하였고, 이를 바탕으로 연구 및 구현에 들어가게 되었다.

표절한 문서를 검출하는 데에는 많은 방법들이 존재하고 있으며, 대표적으로는 단어나 문장들을 가지고 통계적 방법을 이용하여 출현 빈도를 측정하는 방법이 많이 쓰이고 있다.

본 논문에서 제안하는 방법은 입력 받은 사본들 중에서 키워드 중심의 Detector들을 추출하여 핵심 탐지기로서의 역할을 담당하고, 이러한 탐지기들을 동적으로 생성하고 관리하여 불필요한 복잡성(Complexity)을 제거하는데 주안점을 두고 있다. 여기서 Core Detector(핵심 검출기)란 표절을 하면서 많이 쓰일 것 같은 문장의 부분들이 원소로 들어가 있는 집합으로서, 복제의 가능성을 측정하는 탐지기로서의 역할을 하게 된다. 실제로 탐지기에 의해서 복제 가능성이 높은 문건으로 판별이 되면 해당 문서에 대한 정밀 비교가 수행되고 유사율을 산출하게 된다. 앞으로의 본론에서는 이러한 핵심 탐지기를 생성하는 과정과 다양한 길이(Length)를 갖는 탐지기 생성 및 관리하는 방법들을 소개 할 것이다.

2. 자연어 표절 검출 소프트웨어

자연어 표절 검출 소프트웨어는 문서의 구조적인 특징을 검사하는 것보다는 통계적인 특징을 추출하여 표절 검사하는 지문 검사법이나 하이브리드 기법을 주로 사용한다. 현재 국외에 개발된 자연어 표절 검출 소프트웨어는 Digital Integrity의 Findsome [1], CaNexus의 EVE2 [2], iParadigms의 Turnitin [3], CFL Software Developments의 CopyCatch [4], WordCHECK Systems의 WordCHECK [5]등이 있으며, 국내에는 교수클럽 [6]과 Clonechecker [7], LOFC(Linear Order Function Call) [8]등이 존재하나 교수클럽만이 형식을 유지하고 있는 실정이다. COPS [9]는 다양한 형식의 문서들을 ASCII문자로 변환하여 문장 단위로 나눈 후 문장을 그룹 지어서 데이터베이스에 저장을 하고 비교하고자 하는 문서가 들어오면 데이터베이스에 저장되어 있는 기존의 문장들과 중복되어 있는지를 검사한다. SCAM [10]은 COPS를 개선하여 문장 단위가 아닌 단어 단위로 나눈 후 그룹을 지어서 데이터베이스에 저장한 후 비교문서가 들어오면 문서에서 사용된 단어의 빈도수를 벡터로 나타내고 벡터들 간의 dot-product를 통해 유사성을 검출하는 방법이다.

3. 동적비교방법

현재까지 자연어의 복제를 탐지하고, 유사도 및 복제도를 측정하기 위해 다양한 알고리즘이 연구되었고 시스템에 적용되어 왔으나 충분한 성공을 거두지는 못하고 있다. 자연어의 경우 프로그램 소스 코드처럼 특별한 구조를 유지하는 경우도 드물고, 약간의 수정만으로도 의미는 같으나 코드 상으로는 전혀 다른 문장으로 인식이 되어 검출이 쉽지만은 않다. 파일 크기가 작고 약간의 문장이 부분 수정되어 복제 되었다면 문장대 문장(Line By Line)의 비교로서 상당부분 복제에 대한 탐지가 가능할 수 있을 것이다. 하지만 파일 크기가 큰 다수의 문서를 모두 문장 대 문장으로 일일이 비교를 한다는 것은 상당히 많은 컴퓨터의 자원과 시간을 필요로 한다는 단점을 가지게 된다. 이런 단점을 극복하기 위해 효과적이고 탐지 가능성이 매우 높은 탐지기들을 활용하여 수많은 문서들을 일일이 문장 대 문장 또는 단어 대 단어로 비교하지 않고 복제 유무를 신속하고 정확하게 탐지 할 수 있는 방안을 모색하게 되었다.
이 장에서는 기존에 있는 시스템의 단점을 극복하고 시간과 비용, 복잡성(complexity)을 효과적으로 줄일 수 있는 새로운 시스템인 동적 비교 방법을 소개 하고자 한다.

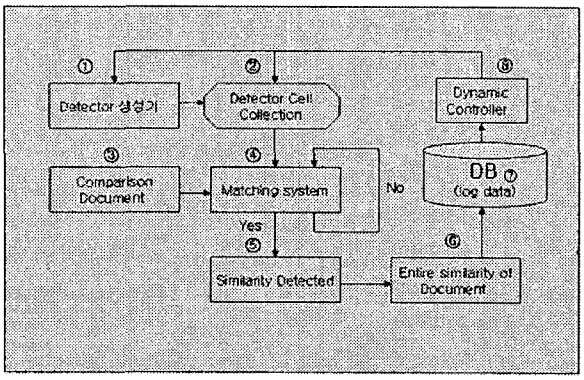

그림 1. 동적 비교 방법의 시스템 구조

4. 동적 비교 방법에 의한 세부 모듈 기능 설명

self Document는 보호해야 될 원본 문서이고 Typical Sample Document 문서는 복제 가능성이 높은 문건이다. 여러 사본들 중 keyword 빈도 수 및 유사성이 높은 문건을 탐지기 생성 비교 문서 자료로 활용한다. 탐지기 생성 비교 문서는 탐지기 생성 및 탐지 효과에 많은 영향을 미치므로 문서를 선택하는데 있어서 신뢰도가 높은 문건을 선택한다. 탐지기 생성을 위해 사본에서 추출한 비교문서(Typical Sample document)를 원본의 Keyword 중심으로 관련된 문장을 filtering한다. Keyword가 포함된 Sample 문서 문장과 원본에 Keyword 포함된 문장이 Keyword 중심으로 좌우로 토근 단위로 비교하면서 일치가 되면 토근의 길이를 확장해나가고 불일치 시 확장을 중단한다.
원본과 사본에서 키워드 중심으로 비교된 다양한 탐지기 Cell들이 생성 되어진다. 실제 사본에서 복제되는 유형을 비교하여 탐지기로 만들었으므로 탐지 가능성을 높였다. 탐지기의 형태는 구, 절, 문장, 문건 등 다양한 크기의 탐지기들이 생성된다. 이렇게 생성된 탐지기들은 탐지기 수집기에 의해서 모아지게 되고 탐지 활동을 하게 된다.

4.1 탐지 서비스의 생성 알고리즘

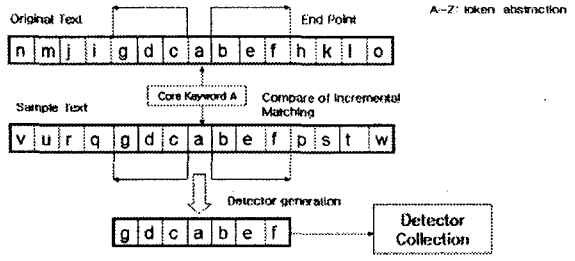

그림 2. 탐지 서비스 생성 알고리즘

A: 원본 문장 B: 사본 문장
A는 n개의 토근으로 이루어져 있음 A: 1.....i .....n
B는 m개의 토근으로 이루어져 있음 B: 1.....j.....m
i 는 비교시 A의 start 토근 위치, j 는 비교시 A의 start 토근 위치

D(A,a): A의 a번째 토근의 데이터 값, D(B,b): B의 b번째 토근의 데이터 값

4.2 탐지 서비스의 특징

① 다수의 문서 중에서 유사도가 높은 문건을 빠르게 탐지 할 수 있다.
② 전체적인 탐지와 세부적인 탐지가 모두 가능하다.
③ 실제 사본에서 매칭 알고리즘을 통하여 탐지기를 생성했기 때문에 매우 유효한 탐지가 가능하다.(실제 사본은 Keyword 분포도를 check해서 복제율이 높은 문서를 채택한다)
④ 핵심 Detector를 실제 사례에서 선형 matching 알고리즘에 의해 다양한 길이의 탐지기를 생성된다.(단어, 구, 절, 문장, 문건등)
⑤ 탐지할 때 비교 횟수를 줄이고 탐지 속도를 향상 시킬 수 있다.
⑥ 탐지 할 때 가능성을 최대한 높은 탐지기를 생성한다.
⑦ 동적으로 탐지기의 생성 및 추가, 삭제하여 능동적으로 환경에 잘 적응할 수 있다.

5. 결론

본 논문에서 제안하는 기법은 표절할 가능성이 가장 높은 정보를 선택적으로 추출하여 가장 합리적이고 신속하게 표절된 문서를 탐지하는데 역점을 두었으며, 불필요한 복잡도(Complexity)를 최소화 하는 점 또한 고려하고 있다. 이러한 탐지기를 바탕으로 수많은 문서들 가운데서 표절된 문서들을 빠르게 찾아냄으로서 효과적으로 표절을 탐지해 낼 수 있는 것이다. 그리고 기존에 있던 Word access pattern, Sentence access pattern이 가지고 있는 긍정적 결함과 부정적 결함을 극복 할 수 있는 하나의 방법으로 입증되었다. 자연어의 특성상 소스코드 표절과 같이 구조적인 접근방법으로는 표절을 탐지하는데 어려움이 있지만 앞으로 좀 더 다양한 시도와 연구가 이뤄져서 해결해야 할 점이란 것은 확실시 되고 있다. 향후 연구과제는 DICOM에서 중요한 Core Detector 생성 과정에서 Keyword를 활용하는 요소가 있는데 Keyword가 분명하지 않은 여러 News Script, 사설, Essay, 연극 대본 등 다양한 문서의 종류에 맞는 새로운 탐지법도 연구가 필요한 사항이고 이에 대한 보완이 이루어 져야 한다고 본다.

참 고 문 헌

1. http://www.findsame.com
2. http://www.caNexus.com
3. http://www.turnitin.com
4. http://www.CopyCatch.freeserve.co.uk
5. http://www.WordCHECKsystems.com
6. http://www.gyosuclub.com
7. http://ropas.kaist.ac.kr/n/clonechecker
8. http://jade.cs.pusan.ac.kr/~hgcho
9. S. Brin, J. Davis, and H. Garcia-Molina, "Copy detection mechanisms digital documents." In Proceeding of the ACM SIGMOD Annual Conference, CA, May 1995
10. Narayanan Shivakumar, HectorGarcia-molina, "Building a Scalable and Accurate Copy Detection Mechanism"