프로그램, 문서에 대한 표절 감정과 미술품, 고문헌등의 지적재산권에 대한 표절 감정 방법의 제안

조 동 욱 * 최 병 갑 **

* : 충북과학대학 정보통신과학과 ** : 목원대학교 컴퓨터공학과

Plagiarism Inspection of S/W Programs, Documents and Plagiarism Inspection Proposal of Intellectual Properties such as Fine Arts and Ancient Literatures

Dong Uk Cho * Byung Kap Choi **

* : Chungbuk Provincial Univ. of Science & Technology ** : Mokwon University

요 약

본 논문에서는 프로그램 소스 코드로부터 표절을 감정하는 기술적 방법론에 대한 고찰과 자연어 형태로 쓰여진 글에 대한 표절 형태 및 이를 감정하기 위한 기술적 방법들에 대해 살펴보고자 한다. 또한 미술품이나 고문헌등에 대한 저작자의 진위 여부 및 표절 감정은 저자뿐만 아니라 소장자의 재산 가치 평가 및 문화재 관리측면에서 대단히 중요한 문제이기 때문에 이를 기술적으로 처리하기 위한 방법론을 제안하고자 한다. 최종적으로 실험에 의해 본 논문의 유용성을 입증코자 한다.

Ⅰ. 서론

IT기술의 발전과 더불어 많은 편리성과 데이터 및 정보의 접근성이 용이해진 반면 이에 대한 역작용등이 사회적 문제로 대두되고 있다[1]~[5].

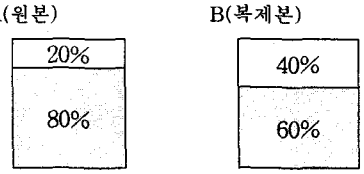
가장 대표적인 역기능이 지적재산권의 표절 행위이다.

본 논문에서는 이를 위해 현재 지적재산권 분야 중에서 가장 법적 분쟁이 많은 분야인 프로그램 소스코드, 자연어로 된 문서, 고문헌이나 예술품의 표절 및 위작 판정에 대한 기술적 방법론의 고찰 및 방법을 새로이 제안하고자 한다. 이중 프로그램 소스코드분야는 표절의 정의 및 감정방법, 표절을 감정할 수 있는 소프트웨어 툴에 대한 분석을 행하고자 한다. 아울러 자연어로 된 문서의 경우 표절의 형태 분석, 방법론, 표절을 감정하기 위한 소프트웨어 툴에 대한 고찰을 행하고자 한다. 끝으로 고문헌과 예술품의 경우 작품의 경계 형태 분석에 의해 위작 여부를 판정할 수 있는 방법을 제안하고자 하며 실험에 의해 본 논문의 유용성을 입증하고자 한다.

2. 프로그램소스코드의 표절 감정

프로그램의 감정이란 컴퓨터 프로그램 및 이와 관련된 전자적 정보등에 대하여 감정인이 전문적 지식과 기술을 활용, 비교 분석하여 복제 여부를 판정해 주는 소견을 말한다[6].

프로그램의 소스코드를 변형하는 방식에는 변수 이름 바꾸기, 명령문의 위치 변경, 주석문의 변경, 동의·중복 연산 사용, 동의 피연산자 사용, 인수 전개등이 있다. 이 같이 변형된 코드에 대해 표절을 감정키 위해서는 프로그램 전문가가 행하는 방법과 표절 감정을 수행하는 소프트웨어툴을 사용하는 방법이 있다. 통상 감정과정은 프로그램 전문가가 행하면서 소프트웨어 툴을 통해 확인하는 방법이 가장 많이 적용된다.



(그림1) 프로그램 복제의 산출 기준

복제도 산출의 기준은 A(예:A의 80%)의 얼마나 많은 부분이 B에 복제되었는가 하는 원본 기준과 B의 얼마나 많은 부분이 복제한 것 인지를 기준(예:B의 60%)으로 하는가 하는 복제본 기준이 될 수 있다. 현재 통상 원본 기준 방식

을 채택하고 있으며 주요 복제 대상으로 핵심(core)코드, 많은 시간/인력 소요 부분, 독창적 아이디어의 구현부분이 될 수 있다. 즉, 일반적인 감정 항목이란 프로그램의 구조, 기능, 소스코드 및 프로그램의 특성에 따른 특정 항목이 되며 감정 항목별 중요도 Wi를 결정하여 최종 복제도를 아래 식(1)과 같이 하여 구한다.

$$d = \sum_{i=1}^{m} Widi \cdot \cdot \cdot \cdot \cdot \cdot \cdot (1)$$

윗식에서 Wi는 항목별 중요도이며, di는 감정 항목별 복제도 계산 결과를 뜻한다. 그러나 현재와 같은 감정 방법은 백분율로 복제도를 나타나는 것에 대한 문제 해결과 감정인이 임의로 설정하는 항목별 중요도 설정에 대한 임의성을 최소화하는 방안이 강구되어야 하는 문제를 내포하고 있다. 아울러 프로그래밍 언어별 복제도 계산방법, 프로그램 규모별 복제도 계산방법, 프로그램 기능/용도별 복제도 계산 방법등에 대한 표준화도 연구과제가 될 수 있다.

2.1 표절 검사 S/W 툴

프로그램 소스코드의 표절을 검사하는 소프트웨어 툴은 크게 구조 연관 검사기법과 지문법이 많이 사용된다. 이중 구조 연관 검사(Structure Related Detection) 방법[7]은 프로그램 소스 코드에서 토큰을 추출하여 나열하면 선형 구조를 이루는 서열이 만들어 지게 되고 프로그램 소스코드는 제어 흐름의 변경이 어렵다는 사실을 이용하여 표절을 검사하는 방법이다. 대표적인 소프트웨어 툴로는 YAP, MOSS, YAP3, Jplag등 [8]이 있으며 아래 <표1>에 이에 대한 소개를 나타내었다.

<표1> 구조 연관 검사기법 소프트웨어 툴

툴이름	방 법	웹 사이트	비 고
YAP	Longest Common Subsequence	http://www.ccsr.cam.ac.uk/~mw263/YAP.html	
YAP3	String Matching Method	http://www.ccsr.cam.ac.uk/~mw263/YAP.html	
Jplag	상 동	http://www.Jplag.de	온라인
MOSS	String Matching Method	http://www.cs.berkeley.edu/~aiken/moss.html	다양한 언어 지원

이에 비해 통계적인 특징을 추출하여 표절을 검사하는 방법이 지문 검사법(Fingerprint Detection)이며, 대표적인 툴이 SIM과 Siff등 이다. SIM은 소스코드를 1바이트씩 토큰들로 만들어서 토큰의 정보와 토큰의 위치들을 배열에다 할당하여 체인 해쉬 기법을 사용하여 비교하는 방식을 채택하고 있고, Siff는 50개의 대표적인 문자들을 추출하여 비교하는 방식을 사용하고 있다. 아래 <표2>에 이에 대한 비교표를 나타내었다.

<표2> 지문검사법을 사용하는 툴에 대한 비교

툴이름	방 법	비 고
Siff	50개의 대표 문자 추출	오프라인
Windiff	비공개	오프라인, GUI환경
SIM	토큰화 시킨후 참조 횟수 비교	오프라인

3. 자연어 표절 감정

3.1 자연어 표절의 형태

자연어의 표절은 통상 아래 <표4>와 같은 형태로 이루어진다[9].

<표4> 자연어 표절의 형태

- ○ 문서의 구조와 구절의 구조가 같은 경우
- ○ 적절한 동의어로 대체한 경우
- ○ 문장의 순서를 바꾼 경우
- ○ 추상적인 개념을 구체적으로 변경하여 서술한 경우
- ○ 잘못된 단어나 어구, 틀린 철자를 그대로 사용하는 경우
- ○ 문장 줄이기
- ○ 문법적 변환
- ○ 문장의 조합

3.2 자연어 표절의 검출 방법

자연어 표절의 검출 방법은 문서의 문맥 흐름 확인, 특정 단어의 사용횟수와 위치 확인, 특정 문장을 추출하여 비교 확인과 통계적인 기법을 적용하는 방법이 있다. 아래 <표5>에 통계적인 기법을 적용하는 방법에 대해 나타내었다.

<표5> 자연어 표절의 검출을 위한 주요한 통계 자료

- ○ 문장의 평균 길이
- ○ 구나 절의 평균 길이
- ○ 핵심 키워드(function words)의 사용 빈도
- ○ 수동태와 능동태의 사용 형태
- ○ 전치사의 사용 빈도
- ○ 단어당 평균 문자수
- ○ 단어의 길이 분포

위의 <표5>의 방법을 이용하여 일반문서 뿐이 아닌 신약성서에서의 바울서신의 위작 여부, 셰스피어와 베이컨의 표절 및 위작 여부등에 대한 판정을 행한 연구도 수행하였다[10].

3.3 자연어 표절 검출 S/W툴

자연어 표절 검출을 사람이 직접 다 한다는 것은 불가능한 일이다. 따라서 이를 행하기 위한 S/W 툴 등이 나와있다. 아래 <표6>에 국내·외 자연어 표절 검출 S/W툴에 대한 소개를 나타내었다.

<표6> 자연어 표절 검출 S/W 툴

상품명	방 법	회사명	웹 사이트
Findsame	비공개	Digital integrity	http://www.digital-integrity.com
EVE2	비공개	CaNexus	http://www.CaNexus.com
Turnitin	비공개	iParadigms	http://www.turnitin.com
Copy Catch	공통어휘 빈도수 검사	CFL s/w Developments	http://www.CoptCatch.freeserve.co.uk
Word CHECK	단어 사용 횟수	WordCHECK Systems	http://www.WordCHECK.systems.com
교수클럽	유사 어절 트리	교수클럽	http://www.gyosnclub.com

4. 예술품의 표절 감정

고문헌 또는 예술품등에 대한 표절 및 저작자 진위 논쟁은 우리나라뿐만 아니라 전세계적으로 많은 논란이 되어 왔다. 그러나 이를 해당분야 전문가에게만 맡길 경우 표절 및 진위 판정에 소요되는 시간과 경비는 경제적으로 막대할 것으로 여겨진다. 따라서 본 논문에서는 패턴 인식 기법을 적용하여 예술품의 진위판정을 행하는 방법론을 제안하고자 한다. 이를 위해 많은 특징 벡터들이 추출되어 표준 패턴과 비교되어야 하는데 본 논문에서는 우선적으로 이 중 저작자의 습작 형태 즉, 곡선과 직선의 처리 형태를 분석하여 이를 히스토그램으로 나타내어 정합을 행하는 방법에 대해 다루고자 한다.

4.1 고예술품의 특징

미술품이나 고문헌의 주요 특징 요소들은 아래 <표7>과 같다.

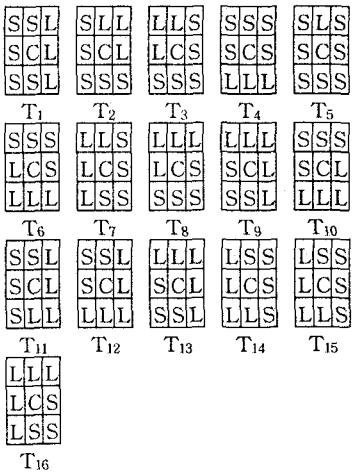
<표7> 미술품이나 고문헌의 주요 특징 요소

대 상	주요 특징 요소
미술품	직선이나 곡선의 처리 형태, 색상 특징, 사용한 재질, 낙관
고문헌	직선이나 곡선의 처리 형태, 서지형태, 사용한 재질, 낙관, 소장인

이중에서 사용한 재질에 대해서는 화학적 분석을 통해 진품 여부를 판정해야 하는데 이를 컴퓨터로 처리를 행할수 없기 때문에 직선이나 곡선의 처리형태, 색상특성, 낙관이나 소장자인에 대한 패턴 인식이 우리가 할수 있는 일로 여겨진다. 따라서 본 논문에서는 우선적으로 전체 진위 판정 시스템중 곡선이나 직선의 처리 형태를 이용하여 진품을 판정하는 방법에 대해 다루고자 하며 차후 색상 특징 분석등과 같은 방법론을 추가로 개발하고자 한다.

4.2 직선이나 곡선의 처리 형태에 의한 진위 여부 판정

저자의 특성에 따라 그림의 둥근 정도나 직선 처리등이 달리 나타나므로 이를 특징 벡터로 선정해야 한다. 이를 위해 아래 (그림1)과 같이 경계 영역의 구성 형태를 16개의 유형으로 구분한다. 이때 전체 영상을 16×16의 부영상으로 나누고 이의 히스토그램 분포를 통해 특징 벡터를 추출한다.



(그림1) 경계 영역의 형태 분석

이때 'S'와 'L'의 계산을 하식 (2), (3)에 의해 구하며 히스토그램의 Y축 눈금 정규화는 식(4)과 같이 정의한다.

$$f_{Large}(X)=\frac{X}{255} \quad \text{-------(2)}$$

여기서 X = |C - L|

$$f_{Small}(X)=\frac{-X+255}{255} \quad \text{------ (3)}$$

여기서 X = |C - S|

$$Y.N=\frac{\sum T_i}{14\times 14} \quad \text{---------- (4)}$$

여기서 i = 1, 2, …, 16이다

이제 각 히스토그램 중에서 히스토그램 누적 분포가 가장 큰 4개의 부분들을 다음과 같은 다항식으로 나타낸다.

(n+1)개의 점에 대하여 n차식의 다항식은

$$(x_0, y_0),(x_1, y_1)\cdots,(x_n, y_n)\cdot\cdot\cdot\cdot\cdot\cdot(5)$$

$$P_n(x)=C_0+C_1X+\cdots+C_nX^n\cdot\cdot\cdot(6)$$

$$P_n(x_k)=y_k\ k=0,1,\cdots,n\cdot\cdot\cdot\cdot\cdot\cdot(7)$$

이다. 차수 n에 대한 Lagrange 다항식은 아래 식(8)을 만족한다.

$$\text{if } i\neq k\ L_k(x_i)=0\ \cdot\cdot\cdot\cdot\cdot\cdot(8)$$
$$else\ L_k(x_i)=1$$

따라서 n차다항식은 하식(9)와 같이 나타내는 것이 가능하게 된다.

$$Pn(x)=\sum_{k=0}^{n} Lk(x)yk=\sum_{k=0}^{n}\left[\prod_{i=0,i\neq k}^{n}\frac{x-x_i}{xk-x_i}\right]yk\cdot\cdot\cdot(9)$$

비매개변수 Hermite 곡선에 대해서는 아래 식(10)과 같이 표현할수 있다.

$$P0,3(x)=y_0L_{0(x)}+y_1L_1(x)+y_2L_2(x)+y_3L_3(x)\cdot\cdot\cdot(10)$$

5. 실험 및 고찰

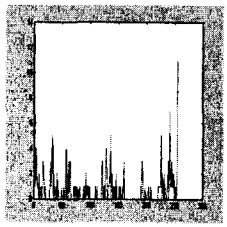
본 논문에서의 시험은 IMB-PC상에서 행하였다. 아래

(그림2)가 원본이고 (그림3),(그림4)가 표절작품이다. 그리고 이에 대해 직선이나 곡선의 처리 형태를 분석한 결과의 예를 히스토그램으로 나타낸 것이 아래 (그림5), (그림6), (그림7)이다. 실험결과에서 알수 있듯이 미술품을 눈으로 보았을시는 구분이 제대로 안되지만 이에 대해 처리 형태 분석을 행한 결과는 원본과 표절작품이 구분이 됨을 확인할 수 있다. 또한 고문헌에 대한 실험예를 (그림8)에 보인다. 이는 청주 고인쇄 박물관에 소장되어 있는 직지인데 이를 모작한 것이 (그림9)이다. 이에 대한 패턴 처리의 결과 예를 (그림10)과 (그림11)에 보인다. 고문헌도 미술품과 마찬가지로 히스토그램의 분석 결과가 차이가 발생함을 확인할 수 있었다. 따라서 본 논문에서 특징 벡터로 선정한 곡선이나 직선의 처리 형태는 대단히 중요한 특징 벡터가 됨을 확인할 수 있었으며 이는 저작의 습작 형태를 분석할 수 있는 유용한 방법임을 알 수 있었다. 향후는 이외에 저자마다 고유한 형태로 주로 사용하고 있는 색상 정보의 분석등도 행하여 중요한 지적재산권인 미술품이나 문헌등의 표절형태를 보다 판별력있게 처리할수 있는 시스템이 되기 위한 확장을 꾀해야 하리라 여겨진다.

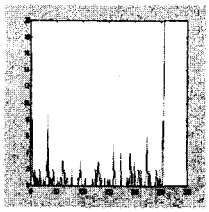
(그림2) 실험미술품(원본)

(그림3) 실험미술품(표절)

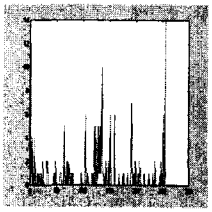
(그림4) 실험미술품(표절)



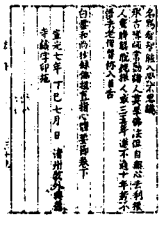
(그림5) 처리형태분석(원본)



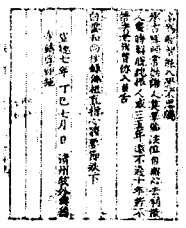
(그림6) 처리형태 분석(표절)



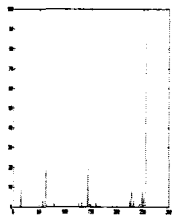
(그림7) 처리 형태 분석 (표절)

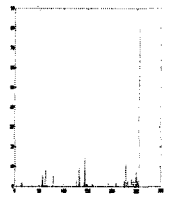


(그림8) 고문헌직지(진품)



(그림9) 고문헌 : 직지 (위작)





(그림10) 처리형태분석 (진품) (그림11)처리형태분석 (위작)

6. 결론

본 논문에서는 지적 재산권의 보호를 위한 프로그램 표절과 다큐먼트 표절에 대한 현재까지의 기술적 방법론에 대한 고찰을 행하였으며, 주요한 지적재산권인 미술품이나 고문헌등과 같은 예술품의 표절 감정을 행하는 방법에 대해 제안하였다. 크게 프로그램 소스코드, 자연어로 된 문서의 표절등에 대해 현재까지의 기술 방법, 현황, 소프트웨어 툴등에 대한 고찰을 행하였으며 예술품의 경우 저작자의 직선이나 곡선 형태 처리를 분석하여 원본과 표절 작품을 판정하는 방법을 제안하였다. 향후 연구과제로는 프로그램 소스코드의 표절을 검출하는 각종의 툴에 대해 여러 각도에서 비교·분석하여 각 도구의 유용성, 제안성, 주요 적용 환경 및 분야, 사용 방법들을 제시하고자 하며 궁극적으로는 표절 검출 도구를 자체 개발하는 것까지 연구를 수행하고자 한다. 또한, 예술품의 표절 검출에 대해서는 색상 정보 분석을 통해 제안한 방법의 유용성을 보완하고 더욱 효율성을 높이기 위한 알고리즘의 추가 개발과 실험 수행등이 행해져야 하리라 여겨진다.

참 고 문 헌

[1] http://www.plagiarism.org

[2] Paul Clough, "Identifying Re-use between the Press Association and Newspapers of the British Press", Dept. of Computer Science, Univ. of Sheffield, Internal Report, June, 2000

[3] Jude Carroll and Jon Appleton, "Plagiarism : A Good Practical Guide," JISC, 2001

[4] 프로그램 감정인 워크샵 자료집, 프로그램심의조정위원회, 2003년, 5월

[5] Joanna Bull, "Technical Review of Plagiariwm Detection s/w Report", Univ. of Luton, JISC, 2001

[6] 이중구, "컴퓨터 프로그램 저작권 보호 제도", 한국멀티미디어학회 워크샵, 2000년, 12월

[7] 조환규, "Genomic Sequence Aligment and its Application to Computing Linear Structure Similarity", 한국생물정보학회, 2002

[8] http://www.cs.usyd.edu.au/~michaelw/YAP.html

[9] 조동욱, 조맹섭, 전병태, "컴퓨터소프트웨어 감정 관련 국내·외 동향 조사 및 분석", 프로그램심의조정위원회 최종연구보고서, 2002

[10] R.D. Lord, Studies in the History of Probaility and Statistics Ⅶ. De Morgan and the Statistical Study of Literature Style, Biometrike, 45, 1958