論文
3-12-2

프로그램 및 자연어 표절 검출을 위한 국내·외 동향 및 감정 S/W 툴의 분석
(Analysis of Domestic·Foreign Trend and Assessment Tools for Programs and Natural Language Plagiarism)

조 동 욱* 신 승 수** 윤 미 희***
(Dong-Uk Cho) (Seung-Soo Shin) (Mi-Hee Yoon)

요 약

컴퓨터 소프트웨어, 디지털 콘텐츠등 디지털 정보 재산권의 보호는 현재 뿐 아니라 향후 국가의 국력을 좌우할 수 있을 정도로 대단히 주요한 과제가 아닐 수 없다. 본 논문에서는 디지털 정보 재산권과 관련된 국내·외 연구 동향과 표절의 감정을 체계적으로 행하기 위한 소프트웨어 툴에 대해 비교·분석을 행하고자 한다. 주요 내용으로는 JISC을 중심으로 한 국외 감정 기관의 사업 동향과 분석 그리고 국내기관이나 단체의 움직임을 살펴보고자 한다. 또한 자연어 표절이나 프로그램 표절을 감정 할 수 있는 S/W툴에 대해 비교·분석을 행하고자 한다.

ABSTRACT

It is very important to protect digital copyright such as computer software, digital contents, and others because national power is influenced at present and in the future. This paper deals with the trend of domestic and foreign researches related to digital copyright and the comparative analysis of software tools for being a systematic judge of the piracy. This paper focuses on the foreign trend of judge business based on JISC and the domestic movement. It also comparatively analyses software tools to judge natural language or program piracy.

* 정회원 : 충북과학대학 정보통신학과 교수
** 정회원 : (주)사이젠텍 연구소장
*** 정회원 : 충북과학대학 컴퓨터정보과학과 교수

논문접수 : 2002. 12. 1.
심사완료 : 2002. 12. 27.

1. 서론

무형적(intangible)이며 비 소모적인 특성을 가지고 있는 소프트웨어나 문헌에 대한 지적재산권 보호는 지식 정보사회에서 대단히 중요한 과제가 아닐 수 없다.

특히 국내의 경우 아직도 소프트웨어의 불법 복제 비율이 50%를 넘어서고 있는 실정이며 이는 인근 아시아와 세계의 불법 복제 비율보다 높은 수준에 이르고 있는 상황이다.

또한, 불법 복제 소프트웨어의 사용 목적이 개인적인 업무를 위한 비중이 60%에 육박하고 있는 실정이어서 이는 차후 상당한 수준으로 원본 소프트웨어의 지적 재산권을 침해하여 경제적으로나 사회적으로 많은 문제점을 야기할 수 있을 것으로 여겨진다.

따라서 불법 복제 방지를 위한 국가적 정책이 수행되어야 하는데 이는 크게 예방정책(Preventative Control)과 억제정책(Deterrent Control)으로 구분되어 질 수 있다.

결론적으로 예방정책과 억제정책이 적절히 조화를 이루어야 하는데 이를 위해서는 프로그램이나 문헌의 불법 복제에 대한 감정이 효과적이고 강력히 수행되어야만 한다.

다시 말해 불법 복제 프로그램이나 문헌에 대한 감정을 효과적으로 수행할 수 있는 기관의 설립과 운영 그리고 표절을 검출할 수 있는 감정 도구의 개발과 활용이 문제 해결의 핵심이라고 할 수 있다. 이를 위해 본 연구에서는 국내·외 감정 기관의 감정 업무 추진 현황을 조사하고 표절 검출 소프트웨어에 대한 비교 분석을 통해 국내 감정 분야의 활성화 및 감정 업무에 활용코자 연구를 수행하였다.

2. 국외의 감정 관련 기관

2.1 영국의 JISC (Joint Information Systems Committee)

- JISC 는 영국, 스코틀랜드, 웨일스와 북아일랜드에서 고등 교육의 질적 향상을 위한 재단법인 FE and HE (Further Education and Higher Education) 산하의 전략적 자문위원회이다.
- 즉, FE 와 HE에서 정보 시스템과 정보 기술의 이용 그리고 창의적 적용을 향상시키기 위한 위원회이다.
- 구체적으로는 네트워크 인프라 스트럭처, ICT(Information and Communication), 정보 서비스, 정보 분야와 관련된 교육 분야 적용이 주된 임무이다.
- JISC 의 조직은 10개의 기구와 3개의 자문위원회, 6개의 실행 위원회로 구성되어 있다.
- JISC 사업중의 하나가 표절(Plagiarism) 검출 및 방지에 대한 사업이다.
- 표절에 대한 "electronic solution" 에 초점이 맞추어져 있다. 즉, 표절 검출 및 표절 정도를 자동으로 계산해주는 S/W 감정도구 개발과 방법론, 비교분석 및 감정도구 이용의 극대화가 주된 사업이 된다.
- 이를 위해 아래 <표 1>과 같이 4개의 프로젝트로 나누어 사업을 수행하고 있다.

<표 1> 전자 표절 검출 및 방지를 위한 JISC 프로젝트의 주된 사업 내용

<Table 1> JISC Project Main Activities for Detecting and Protecting Electronic Piracy

A technical review of free-text plagiarism detection S/W
Source code plagiarism detection S/W
A small size free-text detection S/W
A good practice guide to plagiarism prevention

- 표절 검출 및 방지 대상은 엣세이, 최종보고서, 학위논문과 C++, Java 등으로 작성된 프로그램 등이다.

2.2 미국 및 유럽권

미국 및 유럽권에서 감정관련 활동을 하고 있는 기관은 <표 2>와 같다.

<표 2> 미국 및 유럽권의 활동 기관

<Table 2> Organizations in U.S.A and Europe

국가명	주요 활동 단체명	관련법	참고사이트
미국	BSA(Business S/W Alliance), 소프트웨어 · 정보산업협회 (SIIA : S/W & Information Industry Association)	전자절도방지법 (NET Act : No Electronic Theft Act)	http://www.cybercrime.gov/netsum.htm
독일	저작권 침해 조사 협회 (GVU: Gesellschaftzur Verfalgung von Urheberrechtsverletzungen)	전기통신법 (TeleKommunikationsgesetz)	http://www.gun.de
이태리	저작권 단체 (SIAE)	불법복제방지법 (Anti-Piracy Law)	http://www.siae.it
네덜란드	경제 감찰국 (ECD: Economische Control Dienst). Buma/Sterma(저작권 단체)	저작권법	http://www.burnasternra.nl/bumainternet/homepage.nsf/index.html)
스페인	불법복제 방지 위원회 (Antipiracy Board)	저작권법	http://www.elmundo.es/navegante/200/09/28/escociedad/1001696699.html
호주	AFP (Austrilian Federal Police), ACPR (Austrilian Centre for Policing Research	디지털 의제 (digital agenda)를 위한 저작권법	http://www.copyright.com.au/CopyrightAct.pdf)

2.3 아시아권

주요 아시아 국가에서 감정 관련 활동을 하고 있는 기관은 <표 3>과 같다.

<표 3> 아시아 국가들의 활동 기관

<Table 3> Organizations in Asia

국가명	주요 활동 단체명	관련법	참고 사이트
일본	컴퓨터 소프트웨어 저작권 협회(ACCS:Association of Copyright for ComputerSoftware)	프로바이더 책임법, 부정액세스행위의 금지법	http://www.accsjp.or.jp
홍콩	C&ED	저작권침해금지령 (PCPO:Prevention of Copyright Piracy Ordinance)	http://www.bsa.org/resources/2001-05-21.55.pdf)
중국	불법복제 및 포르노방지청 (NAPP: Office of National Anti-piracy and Pornography) 저작권국(National Copyright Administration)	지적소유권법	http://www.chinapublish.com.cn/e-chinapub/protection/protection01.htm
대만	지적재산권청(Intellectual Property Office), 대만지적소유권연맹(TIPA: Tawiwan Intellectual Property Alliance)	물품검사법, 저작권법	
인도	인도소프트웨어 및 정보서비스 기업연합(NAASCOM: National Association of S/W and Services Companies), 대화형 디지털 S/W 협회(IDSA: Interactive Digital S/W Association)	저작권법	http://www.nasscom.org
태국	지적소유권 및 국제무역중앙 법원 (IP&IT Court: Central Intellectual Property and International Trade Court)	지적소유권법	http://www.iipa.com/rbc/2000/THAILAND.2000.PDF, http://www.iipa.com/2001/2001SPEC301THAILAND.pdf

3. 국내의 감정 관련 기관

3.1 프로그램 심의조정위원회

현재 국내에서는 컴퓨터 프로그램 저작권에 관한 사항의 심의 및 분쟁 조정을 수행하고, 신지식 재산권 제도의 조사/연구를 통해 저작권자의 권리 보호와 공정한 이용 환경의 조성을 위해 프로그램 심의 조정 위원회가 설치·운영되고 있다.

주요활동내역은 아래 <표 4>와 같다.

3.2 한국 S/W 감정 연구회

한국 S/W 감정 연구회는 2001년 11월에 연구회를 창립하여 현재 32명의 전문 감정인이 활동하고 있는 단체로서 연구회원들의 학위별 분류는 아래와 같다.

박사 (72%), 박사수료 (25%), 석사 (3%)

연구회원들의 소속기관 분류는 대학 (52%), 국책연구소(39%), 변호사 등 기타 (9%)이다.

3.3 대학 및 업체 활동

대학으로는 프로그램 표절 검출 S/W 툴을 한국과학 기술원과 부산대에서 개발하였고, 업체에서는 (주)교수클럽이 학생들이 제출한 보고서의 표절 여부를 검사 할 수 있는 S/W 툴을 개발하였다.

국내 개발의 경우 학교에서의 프로그램 복제 검출이나 보고서 표절 검출에 초점을 맞추어 개발되었으며 향후 이에 대한 보완 연구가 지속되어야하리라 여겨지고 있다.

4. 프로그램 표절 검출 S/W 툴

4.1 국외의 경우

프로그램 표절 검출에 쓰일 수 있는 알고리즘은 크게 Attribute Counting 방법과 Structure Metric 방법으로 나뉘어 진다. 이 중 Attribute Counting 방법은 사용된 단어등의 유사성이나 빈도등을 검사하는 방법이고 Structure Metric 방법은 구조적인 방법을 사용하여 정확한 정합이 아닌 토큰 스트링(token string)의 유사성을 계산하는 방법이다. Structure Metric 방법이 Attribute Counting 방법 보다 효율적인 방법으로 평가되고 있으며 Plague, SIM, YAP등의 표절 검출 툴에 사용되고 있다.

<표 4> 프로그램심의위원회 주요활동내역

<Table 4> Main Activities of Program Review Committee

활동 제목	상세 내역
프로그램 저작권 심의/분쟁 조정	프로그램 저작권 관련 정책/기술적 사항을 심의하고 원만한 분쟁 해결을 위한 조정 제도를 통해 프로그램 저작권자의 권리 보호 및 공정한 이용 촉진
프로그램 감정	S/W 디지털 콘텐츠 등 디지털 정보재산권에 대한 분쟁의 공정한 해결을 위해 필요한 감정 기법의 개발 및 연구지원 등을 통하여 감정 업무의 기반을 조성하고 이를 확대함으로써 저작권자 및 이용자 보호
프로그램 등록	프로그램의 창작 사업 및 권리 변동 사항을 분명하게 하여 거래의 안전을 도모하고, 프로그램 홍보 등을 통하여 공시함으로써 중복 투자를 방지함으로써 거래의 안전을 도모하고, 위탁 관리를 통하여 저작권을 체계적으로 관리하고 보호
소프트웨어 지적재산권 조사 연구	국내/외 S/W 지적 재산권 등에 대한 연구를 수행하여 관련 기관 등에 자료를 제공함으로써 지적 재산권 보호
S/W 조건부 일치/프로그램 저작권 위탁 관리	S/W 조건부 일치를 통하여 S/W 사용권자의 권리 보호와 S/W 사장화를 방지함으로써 거래의 안전을 도모하고 위탁관리를 통하여 저작권을 체계적으로 관리하고 보호
S/W 지적 재산권 교육/홍보	S/W의 공정한 이용 마인드의 확산을 위하여 대국민적 교육 및 홍보활동을 전개함으로써 S/W 지적재산권 보호와 관련사업 육성에 기여
S/W 불법 복제 점검 활동 지원	S/W 불법 복제 점검 및 단속 활동에 대한 지원을 통해 불법 복제율을 감소시킴으로써 정품 S/W 사용 정착 유도

아래 <표 5>에 대표적인 표절 검출 S/W 툴에 대한 소개를 나타내었다.

<표 5> 프로그램 표절 검출 소프트웨어

<Table 5> Software for Detecting Program Piracy

시스템	검사대상	개발국	웹 사이트
SIM	소스코드	네덜란드	http://www.few.vu.nl/~dick/sim.html
Dup	소스코드	미국	http://glimpse.arizona.edu/javadup.html
Plague	소스코드	미국	
YAP	소스코드	미국	http://www.ccsr.cam.ac.uk/~mw263/YAP.html
YAP3	소스코드	미국	
MOSS	소스코드	미국	http://www.cs.berkeley.edu/~aikem/moss.html

4.2 국내의 경우

국내의 경우 KAIST의 clonechecker와 부산대의 LOFC가 있다. 아래 <표 6>에 이에 대한 요약표를 나타내었다.

<표 6> 국내 표절 검출 소프트웨어의 비교

<Table 6> Comparison with Domestic Piracy Detection Software

표절 검출 도구	검사대상	검사 방법	측정 가능한 프로그램 언어	웹사이트
Clonechecker	소스코드	프로그램 전체에서 공통된 구문트리 (abstract syntax tree) 가차지하는 비율로서 유사도를 측정	C, Java, Scheme, nML	http://ropas.kaist.ac.kr/n/clonechecker
LOFC	소스코드	키워드를 FUNCTION CALL의 순서에 따라 나열하고 이를 유전자 탐색을 위한 local alignment 문제로 귀착시켜 유사도를 측정함	C, C++, Java, ML	http://jade.cs.pusan.ac.kr/~hgcho

5. 자연어 표절 검출 S/W

자연어 표절의 형태는 아래 <표 7>과 같으며 표절을 검출키 위해 통상 통계적 기법을 사용한다. 이때 사용되는 통계 자료는 아래 <표 8>과 같다.

<표 7> 자연어 표절의 형태

<Table 7> Language Piracy Types

Copying directly form the source
Rewording a sentence (paraphrasing) • Using appropriate synonyms • Changing the sentence type • Changing the order of a sentence • Reducing a clause to a phrase • Changing the part-of-speech • Making abstract ideas more concrete • text summarization - Sentence reduction - Sentence combination - Syntactic transformation - Lexical paraphrasing - Generalisation / specification - Sentence reordering

<표 8> 자연어 표절의 검출을 위한 주요한 통계 자료

<Table 8> Statistics for Detecting Natural Language Piracy

• the average length of sentences (words) • the average length of paragraphs (sentences) • the use of passive voice (expressed as a percentage) • the number of preposition as a percentage of the total number of words • the frequency of "function words" used in each text

통상 통계적 방법을 이용한 클러스터링 방법, 지식 기반 방법을 이용한 클러스터링 방법 그리고 위의 두 방법을 결합한 하이브리드 방법 등이 있다.

5.1 국외의 경우

자연어 표절 검출을 위한 S/W 툴에 대해 아래 <표 9>에 그리고 이에 대한 비교를 <표 10>에 나타내었다.

<표 9> 자연어 표절 검출 소프트웨어

<Table 9> Software for Detecting Natural Language Piracy

시스템	개발 회사명	웹 사이트
Findsame	Digital Integrity	http://www.findsame.com
EVE2	CaNexus	http://www.CaNexus.com
Turnitin	iParadigms	http://www.turnitin.com
CopyCatch	CFL S/W Developments	http://www.CopyCatch.freeserve.co.uk
WordCHECK	WordCHECK Systems	http://www.WordCHECKsystems.com

<표 10> 국외 자연어 표절 검출 소프트웨어에 대한 분석 및 비교

<Table 10> Analysis and Comparison of Foreign Software for Detecting Natural Language Piracy

		Findsame	Eve2	Turnitin	CopyCatch	WordCHECK1
Company	Company	Digital Integrity	CaNexus	iParadigms	CFL Softward Development	WordCHECKsystems
	URL	http://www.findsame.com	http://www.caNexus.com	http://www.turnitin.com	http://www.CopyCatch.freeserve.co.uk	http://www.WordCHECKsystems.com
Delivery Method	Type of system	Web content search system	Compiled local executable, that content searches the Internet	Web-based content search system(of user uploader files and Internet content)	Compiled local content searcher	Compiled local constent searcher
Price	Cost for universities	Free demo	$399 per institution(multi-site), or $19.99 per user	Site license $4000 for unlimited reports	£2000 per institution one-off(negotiable) (upgrades extra)	Basic package for single user is $95(academic price). This archives 1000 documents. additional units are then charged 2,000 units@$295 5,000 units@$895 10,000 units@ $1,495
	Cost for one user	No information	$19.99 one off payment	One year subscription is $100	One off payment of £250	One off payment of $95
Technical	Operating environment	Web-based	PC(windows only)	Web-based	PC(Windows only)	PC(Windows only)
	Ease of mass distribution	Requires browser and Internet connection	Cannot be installed on network for multiple users	Requires browser and Internet connection	Can be installed on network for multiple users	Can be installed on network for multiple users
	Turnaround speed	Instant, dependant on server up-time and Internet traffic	Instant, but local, dependant on processor speed and Internet traffic	24Hrs	Instant, dependant on processor speed	Instant, dependant on processor speed
	Installation engine	N/A	Reliable installer engine	N/A	No installation routine, a manual file transfer	Reliable installer engine
	Reliability	★★	★★★	★★	★★★★★	★★★★★
	Suitability for mass distribution	★★★★★	★★	★★★★★	★★★	★★★
	Stability of vendor	★★★	★★★	★★★★	★	★★★
	Speed of response	★★★	★★★	★★★★	★★★★★	★★★★★
	Technical support	N/A	★★★★	★★	★★★	★

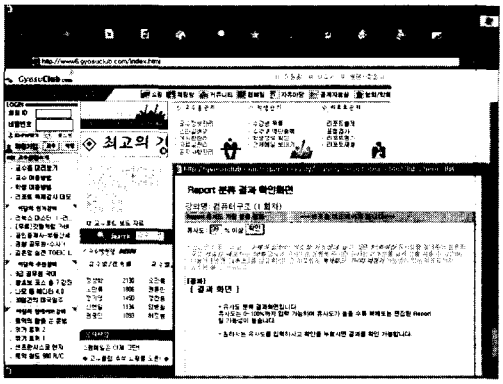
5.2 국내의 경우

국내의 경우 (주) 교수 클럽이 학생들의 레포트 표절 검출을 위해 개발하였다. 아래 <표 11>에 이의 웹사이트 소개를 그리고 [그림 1]에 하나의 예를 나타내었다.

<표 11> 자연어 표절 검출 소프트웨어

<Table 11> Software for Detecting Natural Language Piracy

시스템	개발회사명	웹 사이트
교수클럽	교수클럽	http://www.gyosuclub.com



[그림 1] 레포트 표절 검출의 예

[Fig. 1] Example of Detecting Report Piracy

6. 결론

본 연구에서는 컴퓨터 소프트웨어 감정 관련 국내·외 동향 조사 및 분석에 대한 연구를 수행하였으며 아래와 같은 내용을 조사, 분석하여 수록하였다.

- 국외 감정 기관의 사업 실적 및 동향
  - JISC 의 경우
  - 미국·유럽의 경우
  - 아시아권 국가의 경우
- 국내 감정 기관의 사업 실적 및 동향
  - 프로그램 심의 조정 위원회
  - 한국 소프트웨어 감정 연구회
  - 한국 과학 기술원, 부산대, 교수클럽
- 국외 프로그램 표절 검출 S/W 의 비교·분석
- 국내 프로그램 표절 검출 S/W 의 비교·분석
- 국외 자연어 표절 검출 S/W 의 비교·분석
- 국내 자연어 표절 검출 S/W 의 비교·분석

끝으로 본 연구성에 물심양면으로 지원 해준 프로그램 심의 조정 위원회에게 감사하는 바이다.

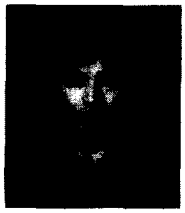
※ 참고문헌

[1] http://www.jisc.ac.uk/jciel/plagiarism
[2] http://cise.sbu.ac.uk/resources.html
[3] http;//www.plagiarism.org
[4] http://www.integriguard.com
[5] http://www.paperbin.com
[6] http://www.canexus.com/eve/abouteve.html
[7] http://www.copycatch.freeserve.co.uk
[8] http://www.plagiarism.com
[9] http://wordchecksystems.com
[10] http://www.ccsr.cam.ac.uk/~mw263/YAP.html
[11] http://www.diglib.stanford.edu
[12] http://www.fewvu.nl/~dick/sim.html
[13] http://glimpse.arizona.edu/javadup.html
[14] http://ftp.cs.berkeley.edu/~aiken/moss.html
[15] http://www.ipd.ira.uka.de:2222/
[16] http://socrates.cs.man.ac.uk/~ajw/pd.html
[17] http://ropas.kaist.ac.kr/n/clonechecker
[18] http://jade.cs.pusan.ac.kr/~hgcho
[19] http://www.findsame.com
[20] http://www.turnitin.com
[21] http://www.gyosuclub.com

조 동 욱

1983년 2월 한양대학교 공과 대학 전자공학과 졸업(공학사)
1985년 8월 한양대학교 대학원 전자공학과 졸업(공학석사)
1989년 2월 한양대학교 대학원 전자통신공학과 졸업(공학박사)
1982년 - 1983년 (주)신도리코 장학생 겸 기술연구소 연구원
1989년 - 1991년 동양공업전문대학 전자통신공학과 조교수
1991년 - 2000년 : 서원대학교 정보통신공학과 부교수
2000년 - 현재 충북과학대학 정보통신공학과 교수
1999년 미국 Oregon state University 교환교수
1996년 한국통신학회, 한국통신학회 충북지부 학술상 수상
1997년 한국통신학회 공로상 수상
1999년 한국통신학회, 한국통신학회 충북지부 학술상 수상
2001년 충청북도지사 표창
2001년 한국정보처리학회 우수논문상 수상
2001년 한국콘텐츠학회 학술상 수상

신 승 수



1984년 3월 - 1988년 2월 충북대학교 수학과졸업(이학사)
1988년 3월 - 1993년 2월 충북대학교 대학원 수학과 졸업(이학석사)
1997년 3월 - 2001년 2월 충북대학교 대학원 수학과 졸업(이학박사)
2000년 3월 - 2002년 9월 (주)시그마정보기술 연구소장
2002년 10월 - 현재 (주)사이젠텍 연구소장
2001년 12월 - 현재 (사)한국콘텐츠학회 총무이사

윤 미 희

1990년 숙명여대 전자계산학과(학사)
1992년 숙명여대 전자계산학과(석사)
1999년 숙명여대 전자계산학과(박사)
2000년 - 현재 충북과학대학 컴퓨터정보과학과 교수