Journal of Digital Contents Society Vol. 7 No. 3 Sept. 2006(pp. 199-205)

디지털 문서 콘텐츠 보호를 위한 문서 복제 탐지 시스템에 관한 연구

김헌*

요 약

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 좀 더 효율적인 방법론과 객관적이고 시스템적인 접근이 필요하다고 본다.

또한 불법적인 지적재산권 침해에 대응한 관리 및 탐지 기술이 더욱 중요해졌음을 의미한다. 본 논문에서는 소중한 지적재산권을 효과적으로 관리 및 탐지하는 기술과 이론을 제시하고자 한다. 또한 기존 DRM 솔루션들이 가지고 있는 장단점들을 분석하여 좀 더 효율적인 디지털 콘텐츠 관리 및 탐지 시스템을 제안하게 되었다.

A Study on the Document Copy Detection System for Protection of Digital Document Contents

Kim Heon*

Abstract

Due to easy access to information in our digital society, there are many cases of illegal counterfeiting and usage of personal information. Producing information with investment and effort is important indeed, but managing and protecting information is becoming a furthermore important issue. This is to promote a new detecting theory and solution for cases of intellectual property violations and plagiarizing digital contents.

Keywords : DRM, Protection, intellectual property right, plagiarizing detection

1. 서론

빠른 속도로 변화하는 시대에 지식과 기술은 기업의 흥망성쇠를 결정짓는 중요한 척도가 되었다. 지금은 장소에 구분 없이 누구나 손쉽게 인터넷 사용을 할 수 있으며 디지털 문서를 다운로드 받을 수 있다.

사용자가 방대한 정보를 편리하게 이용할 수 있는 이면에 디지털 콘텐츠는 원본과 복사본의 차이 없기 때문에 저작권의 불법 사용과 무단 복제의 확산을 막기 위한 디지털 저작권 보호의 중요성이 증대되고 있다.

표절한 문서를 검출하는 데에는 많은 방법들이 존재 하고 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. 앞으

※ 제일저자(First Author) : 김헌
접수일:2006년07월02일, 심사완료:2006년09월04일
* 한신대학교 교양전산학과
heunyong@hs.ac.kr

로의 본론에서는 이러한 핵심 탐지기를 생성하는 과정과 다양한 길이(Length)를 갖는 탐지기 생성 및 관리하는 방법들을 소개 할 것이다.

또한 온라인 및 오프라인 환경에서 모든 콘텐츠 유형에 적용 가능하고 동적인 저작권 관리와 실시간 감시 및 추적이 가능한 디지털 저작권 보호와 디지털 콘텐츠를 무단 복제하거나 불법으로 사용하는 시도를 줄일 수 있는 멀티 에이전트 기반의 DRM 시스템을 제안한다[1]-[3].

2. DRM 기술 개발 동향

2.1 DRM의 개요

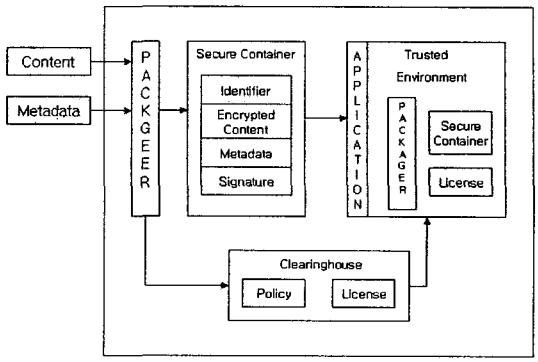
DRM(digital rights management)은 다양한 멀티미디어 콘텐츠의 유통에 있어서 저작권을 보호하고 관리할 수 있는 기술이다. 멀티미디어 콘텐츠의 서비스에 대해서는 서비스 주체별로 다양한 정책이 존재하며, 따라서, DRM도 다양한 형태로 존재할 수 있고 서비스에 적용될 수 있다. DRM의 핵심적 요소기술은 콘텐츠의 보호를 위한 암·복호화 기술과 사용규칙 제어기술, 과금 결제를 위한 기술의 3가지로 분류해 볼 수 있다[4][5].

2.2 디지털콘텐츠 보호 기술

DRM은 관련된 요소 기술이 총체적으로 결합된 솔루션으로서 최근 여러 디지털콘텐츠 보호 유통 기술의 대안으로 제기되고 있다.

DRM 핵심기술은 디지털콘텐츠의 모든 유통 흐름 속에서 지속적으로 적용되어 디지털콘텐츠의 저작권 보호 및 올바른 거래. 분배. 사용이 이루어져야 할 뿐만 아니라 유통 시장의 각 주체들이 다루기 쉽고 일관된 방법으로 기술을 사용 및 적용 가능해야한다

DRM 시스템은 단일한 하나의 시스템으로 구성되어 있다기보다는 디지털콘텐츠 유통의 전반적인 기능과 플로우가 복합된 아키텍처의 형태를 보인다. DRM은 콘텐츠를 메타데이터와 함께 배포 가능한 단위로 패키징하는 "Packager"와 이렇게 배포된 콘텐츠를 사용하고자 하는 사용자의 플랫폼에서 콘텐츠의 이용권한을 통제하고 관리하는 역할인 "DRM Controller", 콘텐츠에 대한 배포 정책 및 라이센스를 발급 관리하는 "Clearinghouse"로 크게 구분할 수 있다[13].



(그림 1) DRM 아키텍처

3. 디지털 콘텐츠 저작권 관리 및 탐지 시스템

제안한 시스템은 서버 측에 있는 센터와 클라이언트 측에 있는 에이전트간에 실시간으로 네트워크를 통하여 디지털 저작권에 대한 보호 및 감시기능이 수행된다. 센터는 외부로부터 저작권 보호 대상 및 보호 조건을 입력받기 위한 외부 인터페이스와, 에이전트와의 통신을 위한 네트워크 모듈과 디지털 저작권 감시와 보호를 위하여 클라이언트측에 에이전트를 파견하고 에이전트로부터 보고 받은 감시결과를 처리하는 관리 모듈과 불법 복사본을 탐지하기 하기 위하여 탐지기 생성 및 탐지기 관리 그리고 탐지활동을 하는 탐지 모듈로 구성된다[7]-[9].

3.1 DICOM(Dynamically Intelligent Content Management) 전체적인 기능과 구조 설계

DICOM의 기본적 역할은 저작권 보호 및 감시를 위하여 파견되는 에이전트들에게 보호 및 감시를 위한 미션을 실시간으로 주는 것과 불법 Copy문서를 탐지하는 것이다.

DICOM은 두 개의 멀티 모듈로 구성되어 있어 사전 및 사후 지적 재산권 보호를 위한 시스템이다.

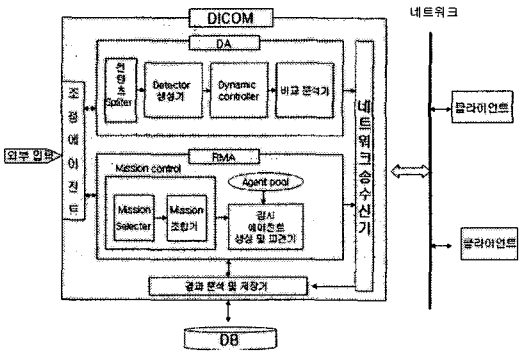
DICOM 내부 모듈의 세부기능은 동적으로 디지털 저작권을 보호 및 탐지할 수 있도록 아래와 같은 기능으로 구성되어 있다(그림 2).

(1)RMA(RightsManagementAgent)모듈설명

- 미션선택기: 보호 대상 자원의 종류와 다양한 보호 조건을 선택할 수 있게 하는 기능
- 미션조합기: 보호를 위한 미션이 동시에 여러 가지 있을 경우 선택된 미션들을 조합하는 기능
- 감시 에이전트 생성 및 파견기: 미션이 설정되었을 때 이 미션을 수행할 에이전트를 생성하고, 보호대상 컴퓨터시스템에 파견하는 기능으로 미션수행을 위한 에이전트 선택을 위하여 에이전트 풀(Pool)을 이용

(2)탐지모듈설명

- 컨텐츠 spliter: 비교해야 될 문서들을 각각sentence 단위로 분할한다.
- Detector 생성기: 탐지기 생성 알고리즘을 통해서 탐지기들을 생성한다.
- Dynamic controller: 생성된 탐지기들을 능동적으로 관리한다.
- 비교 분석기: 탐지기들에 의해서 사본 문서들의 유사율을 실시간으로 산출한다.



(그림 2) DJCOM의 구조

(3) 클라이언트(에이전트) 프로그램의 역할

에이전트는 보호 대상 컴퓨터 시스템에 파견되어 Center로부터 부여된 미션을 수행한다. 감시 에이전트가 수행한 감시 결과를 특정 패턴에 따라서 filtering 하고 filtering된 결과를 온라인 상에서는 센터에 보고하고 오프라인 상에서는 임의의 장소에 감시 결과를 저장한다.

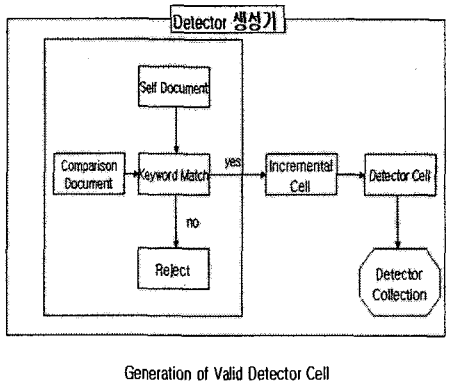
감시 에이전트 내부 모듈의 세부 기능은 네 개의 기능으로 구성되어 있다.

- 임무와 명령 분석기: 부여된 미션을 수행하기 위한 작업을 분석하고 결정하는 기능
- 컴포넌트 구성기: 미션 수행을 위한 작업에 필요한 컴포넌트를 선택하여 구성하는 기능
- 감시 및 보호 수행기: 선택된 컴포넌트를 가동시켜 감시 및 보호를 수행하는 기능
- 감시결과 분석기: 감시 및 보호 결과를 분석하여 센터에 보고 하는 기능

3.2 탐지 Agent 구성 및 기능별 모듈 설명

(1)Detector생성기

Self Document는 보호해야 될 원본 문서이고 Comparison Document 문서는 복제 가능성이 높은 문건이다. 여러 사본들 중 keyword 빈도수 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(Sample document)를 원본의 Keyword 중심으로 관련된 문장을 filtering한다.

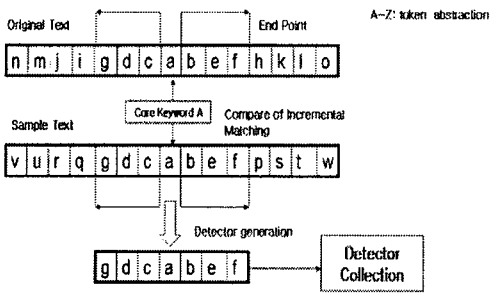


(그림 3) Detector 생성기 구조

원본과 사본에서 키워드 중심으로 비교된 다양한 탐지기 Cell들이 생성되어진다. 실제 사본에서 복제되는 유형을 비교하여 탐지기로 만들었으므로 탐지 가능성을 높였다. 탐지기의 형태

는 구, 절, 문장, 문구 등 다양한 크기의 탐지기들이 생성된다. 이렇게 생성된 탐지기들은 탐지기 수집기에 의해서 모아지게 되고 탐지 활동을 하게 된다 (그림3).

(2)탐지서비스의생성알고리즘



(그림 4) Detector Collection 생성

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 비교 시 A의 start 토근 위치
D(A,a): A의 a번째 토근의 데이터 값, D(B,b): B의 b번째 토근의 데이터 값
If D(A,a) = D(B,b) : a= a-- , b= b-- (until-> a,b >= 1)
Else stop
$\Delta\alpha$= i-a
If D(A,a) = D(B,b) : a= a++, b= b++ (until-> a<=n, b <= m)
Else stop
$\Delta\beta$= a-i
C: Core detector C = Concatenate { D(A, i-$\Delta\alpha$),⋯,D(A, i+ $\Delta\beta$)}

(3) Core Detector 구성 요소

Core Detector를 생성하는데 기준이 되는 요소는 크게 3가지로 구분 되어 진다.

Core Detector
= K(Ct) ∪ R(Ct) ∪ F(Ct)

- K(Ct)는 Keyword와 Typical Sample Document와의 비교에 의해서 생성된 가중치가 높은 Core Detector의 원소들이다.
- R(Ct)는 Relation word와 Typical Sample document와의 비교에 의해 생성된 Core Detector의 원소들이다. Relation word는 Database에 저장 되어 지는 경험에 의한 관계언어들의 집합이다.
- F(Ct)는 입력 받은 문서에서의 단어 빈도수(Word Frequency)를 측정하여, 빈도가 높은 단어(Word)와 Typical Sample document와의 비교를 통해 생성되는 Core Detector의 원소들이다.

3.3 탐지 Agent의 구조적 접근 방법

탐지 에이전트에서 Core detector들을 생성하고 관리하여 실제 탐지 활동에 사용된다. 핵심 키워드와 실제 복사 유형과 매칭 시켜서 생성한 코어 디텍터들은 매우 중요한 의미를 포함하고 있다. 최소한 어느 한 문구, 문장이나 단락에서 중요한 위치를 차지하고 있으며 코어 디텍터들은 주제와 매우 연관성이 깊은 의미적 요소를 가지고 있다.

코어디텍터들의 조합으로 저자가 목표하고 추구하는 뜻이 함축되어져 있다.

다시 말해 이들의 조합에 의해 전체 글들 중에서 가장 핵심적인 내용들을 표현하고 있다고 해야 할 것이다.

이런 의미에서 코어 디텍터간의 관계성이 글의 내용을 결정짓는 중요한 요소가 된다.

실제 복제 유무를 판단하는 기준에서 핵심적인 내용의 포함 여부가 중요한 변수로 작용된다. 구조적으로 복제 여부를 판단하는 요소로서 의미적인 부분이 추가가 되어진다. 의미적으로 파악하는 것은 실제로 많은 어려움이 있으며 문장의 순서와 위치를 바꾸고 여러 수식어로 다른 문장처럼 꾸미면 단순비교를 통해서는 복제가 아닌 다른 문장으로 단정 지을 수 있다.

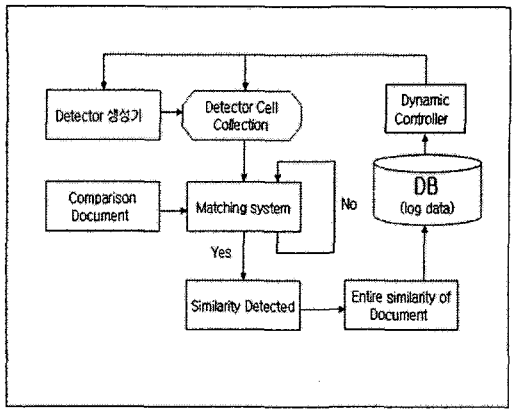
구조적 접근에서 제시하는 방법은 Core Detector간에 관계성을 파악하여 의미적으로 일정한 연결성을 갖는 구조를 파악하여 사전 지식으로 활용하는 방법이다. 하나의 단락에서 몇 개의 Core Detector들이 필수적으로 포함되어야 지만 그 의미를 제대로 전달 할 수 있다면 복제

를 시도하는 사람들도 그러한 문구들은 포함을 시킬 수밖에 없을 것이다. 다른 문구와 문장을 추가하고 원본에 사용된 어휘들을 변경시켜 전체적인 문맥을 바꾼다 하더라도 핵심적 내용은 바뀌지 않게 복제를 한다. 이런 점을 고려하여 관계성을 갖는 핵심 탐지기들이 일정한 범위 안에 포함되어 있는지, 어느 정도 포함되는지 등 의미적으로 영향을 줄 수 있는 요소들을 정의하여 중요한 부분을 복제 하였는지를 판단하는 기준으로 사용 할 수 있도록 하였다.

3.4 탐지 Agent 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, 문구 등)
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 [6][10][11].

(그림 5)는 탐지 에이전트의 전체적인 시스템 설계도 모습을 보여주고 있다.



(그림 5) 전체적인 시스템 설계도

4. 실험 및 평가

● **탐지에이전트실험결과**

실험은 Word access pattern, Sentence access pattern, Core Detector access pattern을 각각 테스트하였다.

100개의 Sample 문서를 테스트한 내용의 결과 값을 정리하여 위에 있는 <표 1>로 나타내었다. <표 1>의 내용을 보면 속도, 복잡성에서 Sentence access pattern이 가장 성능이 가장 우수하다고 할 수 있다. 하지만 실제로 문서 탐지의 효율성과 적합성을 나타내는 Matching rate과 복제 문서 탐지율에서는 가장 낮은 성능을 나타내고 있다. Word access pattern은 Matching rate과 복서 문서 탐지율에서 좋은 성능을 나타내고 있지만 복잡도가 가장 높게 나오고 있다.

논문에서 제안한 Core Detector access pattern은 속도, 복잡도, Matching rate에서 평균 이상의 성능을 나타내고 있으며 특히 복제 문서 탐지율에서 가장 좋은 성능을 가지고 있어서 표절된 문서의 탐지기능으로서 가장 적합하다고 할 수 있다[13].

● **탐지 에이전트 실행 과정**

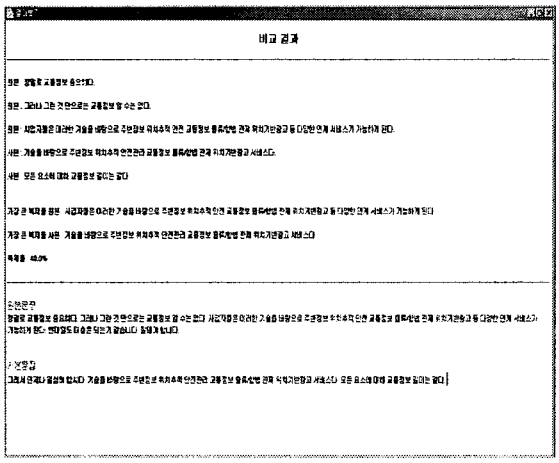
(그림 6)에서 보면 실제 원본과 사본이 비교가 되고 실제 어떤 문장이 일치가 되는지 해당된 문장이 추출이 된다.

또한 구체적으로 전체 문장 중에서 몇 퍼센트의 복제율이 나오는 지 알 수 있으며 원본과 사

본에서 일치되는 문장이 각각 같은 색으로 표시하여 쉽게 눈으로 직접 확인 할 수 있도록 하였다. 모든 문장을 일대일로 순차적으로 비교하는 방식은 비교하는 순서만 변화만 주어도 복제된 부분을 판단 할 수 없지만 탐지 에이전트에서 제안 하는 방식은 불법 복제된 문서에서 Core Detector의 빈도수를 조사하여 복제 유무의 정도를 파악한 후 해당 탐지기 중심으로 선형으로 비교해 나가면서 핵심적인 내용이 복제되었는지 여부를 신속히 탐지해 낼 수 있도록 하고 있으며 불필요한 복잡성을 최소화 하도록 하였다.

<표 1> Performance evaluation by pattern

	Plagiarism detection rate	Speed	Complexity	Matching rate	Reliability
Word access pattern	26.91%	120	82.55	79.45%	42
Sentence access pattern	35.83%	249	56.6	32.53%	51
Core detector access pattern	43.78%	253	69.76	62.58%	76



(그림 6) 불법복제 문서탐지 프로그램의 비교결과

5. 결론

디지털 문화 혁명으로 예전에 생각 할 수 없었던 초고속 멀티미디어 서비스와 다양한 기능의 고부가가치 산업이 창출되었다.

다양한 정보와 지식이 공유되고 유통되어야 하지만 많은 시간과 노력으로 이룩한 소중한 지적 재산은 보호되어야 하고 정당한 대가를 지불해야 하는 것은 이제 하나의 문화로 자리 잡아가고 있다.

디지털 콘텐츠를 보호만 하는데 그치지 않고 사전 및 사후 관리와 법적인 분쟁에도 증거로서 인정받을 수 있는 멀티에이전트 기반의 탐지 및 관리 시스템을 연구하였다.

본 논문에서 제안하는 탐지 에이전트 시스템은 표절할 가능성이 가장 높은 정보를 선택적으로 추출하여 정확하고 신속하게 표절된 문서를 탐지하는데 역점을 두었으며, 불필요한 Complexity를 최소화 하였다. 이러한 탐지를 바탕으로 수많은 문서들 가운데서 표절된 문서들을 빠르게 찾아냄으로서 효과적으로 표절을 탐지해 내는데 있다.

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 접근방법으로 표절을 탐지하는데 어려움이 있지만 앞으로 이런 관점에 대한 연구와 함께 다양한 디지털 콘텐츠 관리와 보호를 위한 심도 있는 연구가 필요하다.

참고 문헌

[1] Yong H. Lee, Su H. Dong and Dae J. Hwang, "Design and Implementation of Digital Rights Management System of KERIS," Proc. of SCI2001 Conference Orlando, FL, USA, July 22-25,, 2001.

[2] Kyung S. Yi, Yong H. Lee and Dae J. Hwang, "implementation of Digital Rights Management System using Active Resource Protection Agent, "Proc. of SCI2001 Conference, Orlando, FL, USA, July 22-25, 2001.

[3] Sun M. Jung, Young M. Kim and Dae J. Hwang, "Implementation of Digital Rights and Protection Rule Database in an Agent based DRM Solution, " Proc. of SCI2001 Conference, Orlando, FL, USA,

July 22-25, 2001
[4] WP4, The IMPRIMATURE Business Model, Ver. 2.1 IMPRIMATYRE, IMP/I40391B, 1999.
[5] X.Wang, XrML: Extensible rights Markup Language Spec. Ver. 1.3, ContentGuard, 2000.
[6] 전명제 "소스 표절 검사를 위한 DIVIDE AND Conquer 방식의 Local Alignment" 부산대학교 소프트웨어 평가학회 춘계 학술대회 논문집, pp.92-94, 2003년6월
[7] NISO, The Dublin Core Metadata Element Set(ANSI/NISO Z39.85-200x), 2000
[8] Renato Iannella, Open Digital Rights Language(ODRL) Ver. 0.8, White Paper, IPR System Pty. Ltd., 2000
[9] Ross J. Anderson, "Information Hiding - A Survey," Proc. of IEEE, Special Issue on Protection of Multimedia Content, May, 1999
[10] 황미녕, 강은미, 한기덕 "유전체 서열의 정렬 기법을 이용한 소스 코드 표절 검사"
[11] 프로그램심위조정위원회 "컴퓨터 소프트웨어 감정관련 국내외 동향 조사 및 분석
[12] 김헌, 조남필, 황대준,"동적 핵심 검출기를 활용한 효율적인 문서 표절 검출 방법" 한국소프트웨어 감정평가학회 춘계학술발표대회
[13] 한국전산원 "URN 기반의 디지털콘텐츠 보급 확산을 위한 전략 수립" March, 2003

김 헌

2005년 : 성균관대학교 컴퓨터공학과(박사)
2005년 ~ 현재 : 한신대학교 교양전산 초빙교수

관심분야 : 멀티미디어, 디지털 컨텐츠, 문서 및 소스 코드 표절 탐지, 인공지능, Bio-Technology