The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 15, No. 1, pp.77-84, Feb. 28, 2015. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2015.15.1.77
JIIBC 2015-1-10

텍스트 기준점 기반의 저작권 침해 판단 시스템 구현

System Implement to Identify Copyright Infringement Based on the Text Reference Point

최경웅*, 박순철**, 양승원***

Kyung-Ung Choi*, Soon-Cheol Park, Seung-Won Yang*****

요 약 기존 문서 저작물의 저작권 침해 판단 방법은 문서를 처음부터 끝까지 문장 단위로 자른 후, 문장 안에서 6어절 단위로 이동하면서 색인키를 생성하여 비교한다. 그러나 이 방법은 문서의 크기가 클 때 색인키가 대량으로 생산되어 표절 검사의 시간이 길어지는 단점이 있다. 이러한 단점을 제거하기 위하여, 본 논문에서 제안하는 방법은 일정한 크기의 윈도우를 문자 단위로 이동하면서 각 윈도우 내에 있는 가장 큰 어절을 선택하여 특징블록을 색인키로 정하는 것이다. 이 방법은 윈도우를 이동하는 과정에서 중복된 특징 블록을 제거하여 색인키의 수를 획기적으로 줄일 수 있다. 이를 적용한 시스템은, 상대적으로 적게 추출된 특징블록간 유사도를 비교함으로써, 두 문서 저작물사이에서 표절된 저작물의 침해 위치를 정확하고 빠르게 검색할 수 있다.

Abstract Most of the existing methods make the index key with every 6 words in every sentence in a document in order to identify copyright infringement between two documents. However, these methods has the disadvantage to take a long time to inspect the copyright infringement because of the long indexing time for the large-scale document. In this paper, we propose a method to select the longest word (called a feature bock) as an index key in the predetermined-sized window which scans a document character by character. This method can be characterized by removing duplicate blocks in the process of scanning a document, dramatically reducing the number of the index keys. The system with this method can find the copyright infringement positions of two documents very accurately and quickly since relatively small number of blocks are compared.

Key Words : Determining Infringement, Copy protection, Text Reference Point

I. 서 론

인터넷과 정보통신기술의 발전은 웹소설, 모바일 소설 등 개인 창작 문서 저작물의 이용방식을 근본적으로 변화시킴으로써 저작물의 상업적•재산적 가치를 현저하게 증대시키고 있다. 한 웹소설 사이트의 2014년 이용 통계조사에 의하면 한해 동안 웹 소설을 올린 작가수는 6만 7천명, 작품수는 전년대비 115% 증가한 12만 3천건이 등록되었다. 그러나 이렇게 대량으로 생산되는 개인 창작물은 창작자의 저작권법 위반에 대한 이해의 부족으로 기존 저작물의 저작권을 침해하는 사례가 빈번히 발생하고 있다. 2014년 저작권 연차 보고서에 따르면 2013년 불법복제물 시장 규모는 3,728억 원에 달하며 이는 전년대비 22%가 증가한 수치다.[1] 검찰청의 2013년 범죄분석

*정회원, ㈜아워텍
**정회원, 전북대학교
***정회원, 우석대학교(교신저자)
접수일자 : 2015년 1월 9일, 수정완료 : 2015년 2월 9일
게재확정일자 : 2015년 2월 13일

Received: 9 January, 2015 / Revised: 9 February, 2015
Accepted: 13 February, 2015
***Corresponding Author: pinksmup@ourtech.co.kr
Chinbuk National Unversity, Woosuk Unversity, Korea

통계자료에 따르면 저작권법 위반 범죄의 30,251건 중 57.3%인 17,334건이 검거되었다[2]. 이처럼 저작물의 불법 복제 등의 침해로부터 저작권자의 권리 보호를 위하여 저작권법에 근거한 규제 및 처벌의 수위가 강화되고 있어 개인 창작자들의 문서 저작물에 대한 저작권 침해를 사전에 예방할 수 있도록 하여야 한다.

기존 문서 저작물의 저작권 침해[3] 판단 방법은 문서를 처음부터 끝까지 각 문장 단위로 자른 후 문장 안에서 6어절 단위로 쉬프트 하면서 색인키를 생성하여 비교하는데, 이 경우 문서의 크기가 클 경우 색인키가 대량 생산되어 표절 검사의 시간이 길어지는 단점이 있다. 본 논문에서는 개인 창작자들이 생산한 문서 저작물의 저작권 침해를 예방하기 위하여 사전에 진단할 수 있는 시스템을 제시하고 저작물의 침해 위치를 빠르게 검색할 수 있는 알고리즘을 구현하도록 한다.

II. 저작권 침해 판단 시스템

1. 저작권 침해 판단 시스템 구성

본 논문에서 구현한 저작권 침해 진단 시스템은 텍스트 기준점 기반의 저작권 침해 판단 알고리즘을 이용하여 인터넷을 통한 저작권 침해 진단 요구 정보를 받기 위한 웹서비스모듈과, 텍스트 기준점을 자동으로 추출하여 검색엔진에서 색인화 및 검색할 수 있도록 처리하는 침해진단모듈과, 실제로 저작권 문서의 색인 정보 저장과 검색을 처리하는 검색엔진모듈과 사용자 정보, 저작권 메타 정보, 서비스 사용 정보 등 시스템 전반을 관리하기 위한 시스템 관리 DB와, 웹브라우저가 아닌 일반 어플리케이션에서 저작권 침해 진단 시스템의 서비스를 사용하게 도와주는 API라이브러리를 포함하여 구성하고 있다.

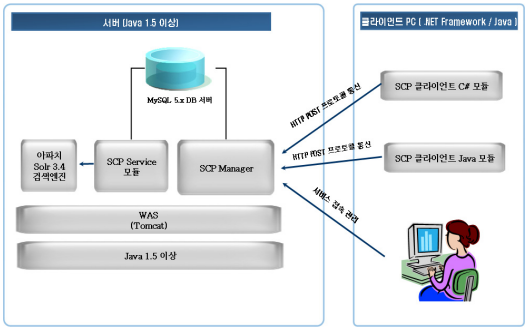

그림 1. 시스템 구성도
Fig. 1. System Diagram

2. 저작권 침해 판단 알고리즘

웹서비스모듈을 이용하여 등록된 문서는 침해진단모듈로 전달된다. 전달된 문서는 시스템에 등록할 경우에는 색인 단계, 시스템에 있는 저작권 문서와 비교할 경우에는 저작권 침해 진단 단계를 수행한다. 먼저, 색인단계 및 침해진단단계는 윈도우 단위의 텍스트 기준점을 추출하는 기능을 수행하게 된다. 텍스트 기준점 방법은 윈도우(W), 기준점(F), 기준점블럭(B)을 이용하여 처리하며 기본 방법은 다음과 같다.

입력되는 문서는 다음과 같이 정의한다.

$$D_1 = E_1, E_2, E_3, E_4 \ldots\ldots E_a \quad (1)$$

수식 (1)의 D_i는 색인을 하기 위한 i번째 문서를 의미한다. E_i는 $E_1 \sim E_i$ 중에서 $_i$번째 어절을 의미한다. 어절은 기본적으로 문서의 공백문자를 기준으로 분리한 것을 의미하며 추가적으로 심볼이나 숫자등도 같이 이용될 수 있다. 문서는 위의 식처럼 $E_1 \sim Ea$까지의 N개의 어절로 이루어진 순차적인 집합으로 정의 될 수 있다.

윈도우(W)는 문서(D_i)에서 기준점을 찾기 위한 순차적인 어절의 부분집합을 의미하며 부분집합의 크기를 윈도우 크기로 정의한다.

$$W_i(s) = \{E_i, E_{i+1}, \ldots\ldots, E_{i+s}\} \quad (2)$$

W_i는 i번째 윈도우를 의미하며 s는 윈도우 크기를 나타낸다. 수식 (2)의 $W_{i(s)}$는 i번째 윈도우에서 윈도우 크기가 s인 부분집합을 의미한다.

표 1. 윈도우 정의
Table 1. Define of Windows

윈도우 번호	윈도우 집합
$W_1(30)$	$\{E_1, E_2, E_3, E_4, \ldots\ldots, E_{27}, E_{28}, E_{29}, E_{30}\}$
$W_2(30)$	$\{E_2, E_3, E_4, E_5, \ldots\ldots, E_{28}, E_{29}, E_{30}, E_{31}\}$
...	...
$W_{70}(30)$	$\{E_{70}, E_{71}, E_{72}, E_{73}, \ldots\ldots, E_{96}, E_{97}, E_{98}, E_{99}\}$
$W_{71}(30)$	$\{E_{71}, E_{72}, E_{73}, E_{74}, \ldots\ldots, E_{97}, E_{98}, E_{99}, E_{100}\}$

예를 들어, $D_i = \{E_1, E_2, E_3, E_4, \ldots\ldots E_{100}\}$, $E_1 \sim E_{100}$까지로 정의된 문서(D_i)가 있다고 가정하고 s가 30인 $W_i(30)$는 표1과 같이 표시할 수 있다. 윈도우(W)가 정

해지면 윈도우마다 기준점(F)을 정하게 된다. 기준점(F)은 윈도우 집합에서 m개의 순차적 어절의 길이 합이 최대인 어절의 순차적 집합을 의미한다. 기준점(F)이 정해지면 기준점을 포함하여 좌우로 k개의 어절을 포함한 순차적 어절의 집합을 기준점블럭(B)으로 정의한다.

$$\mathrm{SUM}_j(m)=\sum_{k=j}^{j+m-1} len(E_k) \tag{3}$$

$$F_i(m)=MAX(\ \mathrm{SUM}_j(m)\ : j=i, i+1, i+2, i+s-m) \tag{4}$$

$$B_{i(k)}=\{E_{j-2}, E_{j-1}, E_j, E_{j+1}, E_{j+2}, ..., E_{j+k}\},\quad k>m \tag{5}$$

수식 (3)의 $\mathrm{SUM}_j(m)$은 j번째 어절부터 m개 어절을 합한 길이를 의미한다. 수식 (4)의 기준점 $F_{i(m)}$는 $MAX(\ )$ 함수를 통하여 수식(2)의 $W_i(s)$에서 수식(3)의 $\mathrm{SUM}_{j(m)}$의 값이 최대인 어절의 순차적 집합을 구한다. 수식 (5)의 기준점블럭 $B_i(k)$는 수식 (4)의 기준점 $F_{i(m)}$을 포함하는 좌우 k 만큼의 크기를 포함하는 순차적 어절의 집합을 의미한다.

표 2. 기준점 블록 선택
Table 2. Choice of Text Reference Point

윈도우 번호	기준점(예시)	기준점블럭(예시)
$W_1(30)$	$F_1(3)=\{E_{10}, E_{11}, E_{12}\}$	$B_1(5)=\{E_5, ..., E_{10}, E_{11}, E_{12}, ..., E_{15}\}$
$W_2(30)$	$F_2(3)=\{E_{10}, E_{11}, E_{12}\}$	$B_2(5)=\{E_5, ..., E_{10}, E_{11}, E_{12}, ..., E_{15}\}$
$W_{70}(30)$	$F_4(3)=\{E_{80}, E_{81}, E_{82}\}$	$B_3(5)=\{E_5, ..., E1E_{11}, E_{12}, ..., E_{15}\}$
.....	.....	.....
$W_{40}(30)$	$F_{40}(3)=\{E_{47}, E_{48}, E_{49}\}$	$B_{40}(5)=\{E_{42}, ..., E_{47}, E_{48}, E_{49}, ..., E_{52}\}$
$W_{40}(30)$	$F_{41}(3)=\{E_{47}, E_{48}, E_{49}\}$	$B_{41}(5)=\{E_{42}, ..., E_{47}, E_{48}, E_{49}, ..., E_{52}\}$
$W_{40}(30)$	$F_{42}(3)=\{E_{47}, E_{48}, E_{49}\}$	$B_{42}(5)=\{E_{42}, ..., E_{47}, E_{48}, E_{49}, ..., E_{52}\}$
.....	.....	.....
$W_{70}(30)$	$F_{70}(3)=\{E_{80}, E_{81}, E_{82}\}$	$B_{70}(5)=\{E_{75}, ..., E_{80}, E_{81}, E_{82}, ..., E_{85}\}$
$W_{71}(30)$	$F_{71}(3)=\{E_{80}, E_{81}, E_{82}\}$	$B_{71}(5)=\{E_{75}, ..., E_{80}, E_{81}, E_{82}, ..., E_{85}\}$

예를 들어, 표 1에서 예로 정의된 $W_i(30)$에서 3개의 어절의 길이 합이 최대인 것, 즉 $F_i(30)$을 기준점으로 하고, 기준점을 포함하여 좌우로 5개, 즉 $B_i(5)$를 기준점블럭으로 정했다면 $F_i(3)$와 $B_i(5)$는 [표2]와 같이 된다.

표 2에서 $F_1(3)$은 $W_1(30)$에서 $\mathrm{SUM}_j(3), j=1,2,...,29$를 표 3처럼 가정 할 경우, 최대인 어절 집합을 $\{E_{10}, E_{11}, E_{12}\}$로 예를 든 것이다. $\mathrm{SUM}_j(3)$의 최대값이 동일하게 존재할 경우 첫 번째 최대값을 선택한다.

표 3. 기준점 블록 추출
Table 3. Extraction of Text Reference Point

$\mathrm{SUM}_j(3)$	수식	어절 길이 합(예시)
$\mathrm{SUM}_1(3)$	$Len(E_1)+Len(E_2)+Len(E_3)$	7
$\mathrm{SUM}_2(3)$	$Len(E_2)+Len(E_3)+Len(E_4)$	7
$\mathrm{SUM}_3(3)$	$Len(E_3)+Len(E_4)+Len(E_5)$	8
.....	.....	.....
$\mathrm{SUM}_9(3)$	$Len(E_9)+Len(E_{10})+Len(E_{11})$	10
$\mathrm{SUM}_{10}(3)$	$Len(E_{10})+Len(E_{11})+Len(E_{12})$	13
$\mathrm{SUM}_{11}(3)$	$Len(E_{11})+Len(E_{12})+Len(E_{13})$	12
.....	.....	.....
$\mathrm{SUM}_{26}(3)$	$Len(E_{26})+Len(E_{27})+Len(E_{28})$	13
$\mathrm{SUM}_{27}(3)$	$Len(E_{27})+Len(E_{28})+Len(E_{29})$	11
$\mathrm{SUM}_{28}(3)$	$Len(E_{28})+Len(E_{29})+Len(E_{30})$	9

$B_1(5)$은 E_{10}을 기준으로 좌우로 5개의 어절을 포함한 것을 예로 든 것이다. 기준점과 기준점 블록 추출이 완료되면 n개의 어절로 이루어진 D_i를 다음과 같이 기준점(F)과 기준점블럭(B)으로 다시 정의 할 수 있다. s는 윈도우 크기, m은 기준점 어절 수, k는 기준점 블록 크기를 의미한다.

$$D_i=(F_1(m), B_1(k)), (F_2(m), B_2(k)), ..., F_{n-s+1}(m), B_{n-s+1}(K)) \tag{6}$$

수식 (6)처럼 기준점과 기준점블럭으로 이루어진 D_i를 구성하고 나면 D_i에서 중복되는 기준점을 제거하고 검색엔진에 색인할 수 있도록 한다. 예를 들어, [표 3]에서 $W_1(30)$, $W_2(30)$, $W_3(30)$, $W_4(30)$가 동일하여 $W_1(30)$ 하나만 선택하고, $W_{40}(30)$, $W_{41}(30)$, $W_{41}(30)$, $W_{42}(30)$가 동일하여 $W_{40}(30)$ 하나만 선택하고, $W_{70}(30)$, $W_{71}(30)$이 동일하여 $W_{70}(30)$ 하나만 선택한다. 위의 [표3]에서 중복을 제거하고 D_i를 표현하면 수식 (7)과 같이 정의 될 수 있게 된다.

$$D_i = \{(F_1(3), B_1(5)), ..., (F_{40}(3), B_{40}(5)), ..., (F_{70}(3), B_{70}(5))\} \quad (7)$$

검색엔진에 중복되지 않는 기준점과 기준점블럭 정보가 색인 완료되면 질의 문서를 기준점 기반으로 검색하여 저작권 침해 진단을 할 수 있다. 색인된 문서는 F, B 앞에 D를 붙이고 질의 문서는 F, B 앞에 Q를 붙여서 수식 (8)과 같은 색인문서 D_i와 수식 (9)와 같은 질의 문서 Q를 정의할 수 있다. 중복되는 기준점은 하나만 사용한다.

$$D_i = \{(DF_1(m), DB_1(k), ..., (DF_{20}(m), DB_{20}(k), ...\} \quad (8)$$

$$Q = \{(QF_1(m), QB_1(K)), ..., (QF_{50}(m), QB_{50}(K)), ...\} \quad (9)$$

질의문서 Q도 역시 기준점과 기준점 블록으로 동일하게 표현할 수 있으며, 윈도우 크기(s) 및 기준점 어절 수(m), 기준점 블록 크기(k)등은 색인 설정과 동일해야 한다. 예를 들어, 질의문서 Q에서 D_i문서의 $DB_{20}(k)$를 복사한 부분이 $QB_{50}(k)$라고 가정하면 $QF_{50}(m)$과 $DF_{20}(m)$의 기준점이 동일하게 되어 질의문서 Q의 $QF_{50}(m)$을 검색하게 되면 기준점이 같은 $DF_{20}(m)$을 찾을 수 있게 된다. 하지만, $QF_{50}(m)$과 같은 기준점은 D_i 문서의 $DF_{20}(m)$ 외에도 여러 개 존재할 수 있다. 이러한 경우 $QF_{50}(m)$과 같은 기준점으로 필터링한 후에 $QB_{50}(k)$ 기준점블록에서 임의로 선택된 검색어를 필터링 된 기준점블럭에 질의한다. 그러면 검색엔진은 검색어와 유사도가 높은 $RB_i(k)$가 결과로 나오게 된다. $RB_i(k)$는 유사도 순위가 i번째인 검색된 기준점블럭을 의미한다. 여기에서 검색 결과 유사도가 임계치 이상인 $RB_i(k)$를 선택하여 $QB_{50}(k)$과 기준점블럭 사이의 유사도를 다시 계산하여 침해 진단을 판단한다. 따라서 기준점을 기반으로 색인하게 되면 검색할 때 모든 기준점을 검색하지 않고 동일한 기준점으로 제한할 수 있어 검색 속도 향상과 침해 위치를 정확히 알 수 있게 된다.

침해진단모듈은 위에서 설명한 윈도우를 이용한 기준점 기반의 색인 정보 추출 및 저작권 침해 진단을 처리한다. 색인 단계는 웹서비스모듈에서 전달 받은 색인 대상 문서의 입력, 윈도우를 이용한 기준점 추출, 색인 정보 선택, 검색엔진 저장 단계로 처리한다.

색인 대상 문서는 기준점을 추출하기 위해 문서를 어절로 분리하며 수식 (10)과 같이 D_i를 어절 단위로 구성한다.

$$D_i = \{E_1, E_2, E_3, E_4, ..., E_n\} \quad (10)$$

어절로 분리하고 나면 윈도우 크기가 s인 $S_i(s)$로 분리한다. 윈도우 크기(s)는 추출되는 기준점 수에 영향을 준다. 윈도우 크기가 클수록 기준점 수가 적어지게 되고, 윈도우 크기가 작을수록 기준점 수가 증가하게 된다. 윈도우 크기가 클 경우 완전 복사한 침해 문서를 찾기는 쉬워도 작은 영역의 부분 복사는 침해 진단을 못하는 확률이 커진다. 반대로, 윈도우 크기가 작을 경우 완전 복사한 침해 문서부터 부분 복사한 침해 진단까지 가능하나 기준점 수가 많이 추출된다. 따라서 윈도우 크기는 부분 복사의 크기를 어느 정도까지 진단할 수 있는가와 시스템이 허용할 수 있는 총 기준점 수를 결정하여 윈도우 크기를 정한다.

$W_i(s)$로 분리하고 나면 윈도우마다 기준점 $F_i(m)$과 기준점블럭 $B_i(k)$를 추출한다. 기준점과 기준점블럭이 추출되고 나면 수식(11)과 같이 D_i가 정의된다.

$$D_i = \begin{Bmatrix}(F_1(m), B_1(k)), (F_2(m), B_2(k)), ..., (F_{n-s+1}(k), \\ B_{n-s+1}(k)\end{Bmatrix} \quad (11)$$

윈도우를 이용한 기준점이 추출[4]되게 되면 중복되는 기준점을 제거하는 색인 정보 선택 단계로 이동한다. 기준점 $F_i(m)$은 윈도우 $W_i(s)$에서 m개 어절 길이의 합이 최대인 것을 기준점으로 결정하기 때문에 윈도우가 한 어절씩 이동하더라도 기준점의 변동은 자주 발생하지 않는다. 중복되는 기준점은 하나만 선택하여 D_i문서를 다시 중복되지 않는 기준점과 기준점블럭으로 구성한다. 예를 들어, D_i문서에서 중복되지 않은 선택된 기준점이 $F_1(m)$, $F_{20}(m)$, $F_{50}(m)$, $F_{80}(m)$ 라고 가정하면 수식(12)처럼 D_i문서를 다시 표현할 수 있다.

$$D_i = \begin{Bmatrix}(F_1(m), B_1(k)), (F_{20}(m), B_{20}(k)), (F_{50}(k), B_{50}(k)), \\ (F_{80}(k), B_{80}(k))\end{Bmatrix} \quad (12)$$

색인 정보가 선택되고 나면 검색엔진에 색인정보를 전달하여 실제로 색인을 진행하는 검색엔진 저장 단계로 이동한다. 기준점 $F_i(m)$은 m개의 어절 집합으로 검색할 때 효율을 높이기 위해 동일한 길이로 변환하여 색인한다. 기준점$F_i(m)$을 수식(13)과 같이 해시함수를 이용하여 동일한 길이로 변환할 수 있다.

$$H_i(m) = hash(F_i(m)) \quad (13)$$

수식 (13)에서 기준점$F_i(m)$에 속한 m개의 어절을 hash() 함수에 입력하면 분리되어 있는 입력된 어절을 모두 하나로 연결하고 난 다음 해시키로 변환하여 반환한다.[5] 예를 들면, 수식 (12)에서 기준점이 $F_1(m)$, $F_{20}(m)$, $F_{50}(m)$, $F_{80}(m)$이 선택된 문서는 검색엔진에 저장하기 위해서 수식(14)처럼 기준점이 해시키로 변환된다

$$D_i = \begin{Bmatrix}(H_1(m), B_1(k)), (H_{20}(m), B20(k)), (H_{50}(m), B_{50}(k)), \\ (H_{80}(m), B_{80}(k))\end{Bmatrix} \quad (14)$$

검색엔진에서는 수식 (14)처럼 기준점 해시키 $H_i(m)$와 기준점블럭 $B_{i(k)}$을 하나의 레코드처럼 저장하며, 기준점 해시키와 기준점블럭을 색인하게 된다. 기준점 해시키는 해시키 값을 색인하게 되며, 기준점블럭은 기준점블럭에 포함된 어절 E_i를 색인하게 된다. 저작권 침해 진단 단계는 웹서비스모듈에서 전달 받은 질의 문서 입력, 윈도우를 이용한 기준점 추출, N개 기준점 선택, 검색어 선택, 검색엔진 질의, 매칭률계산, 결과제공 단계로 처리한다. 웹브라우저 또는 API를 이용하여 웹서비스모듈에 전송된 질의 문서를 웹서비스모듈로부터 전달 받는다. 질의문서 Q도 역시 기준점과 기준점 블록으로 동일하게 표현할 수 있으며, 윈도우 크기(s) 및 기준점 어절 수 (m), 기준점 블록 크기(k)등은 색인 설정과 동일해야 한다. 질의 문서 Q는 기준점을 추출하기 위해 문서를 어절로 분리하며 수식(15)와 같이 Q를 어절 단위로 구성한다.

$$Q = \{E_1, E_2, E_3, E_4, \ldots E_n\} \quad (15)$$

색인단계와 동일하게 $W_i(s)$로 분리하고 나면 윈도우마다 기준점 $F_i(m)$과 기준점블럭 $B_i(k)$를 추출할 수 있고 기준점 $F_i(m)$를 해시키로 변환하면 수식(16)과 같이 Q를 다시 정의할 수 있다.

$$Q = \begin{Bmatrix}(H_1(m), B_1(k)), (H_2(m), B_2(k)), \ldots, (H_{n-s+1}(m) \\ B_{n-s+1}(k))\end{Bmatrix} \quad (16)$$

질의문서 Q가 기준점 해시키와 기준점블럭으로 재정의하고 나면 중복되는 기준점 해시키를 제거하고 한 번에 검색엔진에 질의할 수 있는 N개 기준점 선택 단계로 이동한다. 검색엔진은 기준점 해시키 및 검색어 질의를 한번에 OR 조건으로 질의할 수 있는 최대값이 존재한다. 한 번에 질의할 수 있는 검색엔진 최대값 보다 작거나 같게 N개 기준점을 선택해야 한다. 예를 들어, 질의문서 Q에서 추출된 중복되지 않는 기준점이 1000개이고 검색엔진이 한 번에 최대 100개를 질의할 수 있다면, 질의를 위한 기준점 선택은 최소 1개에서 최대 100개를 지정할 수 있게 된다. 기준점 100개를 지정했다면 최대 10번을 반복해서 검색하면 질의문서 Q의 1000개 기준점에 대해 모두 검색을 할 수 있게 된다.

질의문서 Q에서 한 번에 질의할 수 있는 N개 기준점이 선택되고 나면 선택된 기준점블럭에서 검색어를 선택하는 단계로 이동한다. 기준점 해시키만 가지고 검색엔진에 질의할 수 도 있지만 동일한 기준점 해시키가 여러 개 존재할 경우 검색된 모든 결과에 대해 기준점블럭을 검사해야 한다. 하지만, 검색어를 같이 질의하면 검색어와 색인된 기준점블럭이 유사도가 높은 순으로 정렬할 수 있기 때문에 임계치를 설정하여 임계치 이상일 경우에만 침해라고 판단을 한다면 검색속도를 향상 시킬 수 있다.

검색어를 선택하기 위해 기준점블럭을 하나의 문서라고 가정한다면 tf-idf[6] 가중치를 이용할 수 있게 된다.

$$r_{ie} = f_{ie} * \log(\frac{N}{n_e}) \quad (17)$$

수식 (17)의 r_{ie}는 i번째 기준점블럭 $B_i(k)$에서 어절

e의 tf-idf 가중치를 의미하며, f_{ie}는 i번째 기준점블럭 $B_i(k)$에서 어절 e의 출현빈도를 의미하며, N은 선택된 기준점블럭 수를 의미하고, n_e는 어절 e가 출현한 기준점블럭 수를 의미한다. tf-idf 가중치인 r_{ie}가 높은 N/2개의 어절 E_i를 선택하며 r_{ie}가 동일한 경우에는 어절 E_i의 길이가 큰 것을 먼저 선택한다. 선택된 검색어 어절 E_i가 기준점블럭에 최소한 한 개 이상 포함되는지 확인하고 E_i가 기준점블록에 없다면 기준점블럭에서 최대 길이의 어절을 선택하여 검색어에 추가한다. 이렇게 하면 기준점블럭마다 최소 한 개의 검색어가 포함되며 최대 N개의 검색어를 선택할 수 있다. 질의문서 Q에서 N개 기준점 해시키와 검색어가 선택되고 나면 실제로 검색하는 검색엔진 질의 단계로 이동한다. 검색엔진에 기준점 해시키를 OR 연산자로 검색하여 필터링을 한 후에 검색어를 OR 연산자로 색인된 기준점블럭을 검색하면 유사도가 높은 순으로 검색결과를 가져올 수 있다. 유사도가 높은 순으로 정렬할 수 있기 때문에 임계치를 설정하여 임계치 이상만 침해 진단을 한다면 검색속도를 향상 시킬 수 있다. n개의 검색결과R은 기준점 해시키(H), 기준점블럭 (B) 앞에 R을 붙여 수식 (18)과 같이 표시한다.

$$R=\begin{Bmatrix} RH_1(m), RB_1(k)), RH_2(m), RB_2(k)), ..., RH_n(m), \\ RB_n(k)) \end{Bmatrix} \quad (18)$$

$RH_i(m)$은 유사도 순위가 i번째인 검색된 기준점 해시키를 의미하며, $RB_i(k)$는 유사도 순위가 i번째인 검색된 기준점블럭을 의미한다. $RH_i(m)$은 질의한 기준점 해시키 중에 포함된 값이다. 검색결과 R을 구하고 나면 $RH_i(m)$와 질의문서 $QH_i(m)$이 같은 것을 찾아 $RB_i(k))$와 $RB_i(k)$의 유사도를 구하는 매칭률 계산 단계로 이동한다. 기준점블럭의 유사도

SIM(RB_i(k), QB_i(k))은 수식 (19)와 같이 계산될 수 있다.

$$\mathrm{SIM}\left(RB_i(k), QB_i(k)\right) = |RB_i(k) \cap QB_i(k)| \quad (19)$$

여기서 $|QB_i(k)|$는 질의문서 $QB_i(k)$의 기준점블럭에 포함된 어절 수를 의미하고, $|RB_i(k) \cap QB_i(k)|$은 질의문서 $QB_i(k)$와 $RB_i(k)$의 기준점블럭 교집합을 구한 어절의 수를 의미한다. SIM$RB_i(k)$, $|RB_i(k) \cap QB_i(k)|$ 값이 임계치 이상이면 최종적으로 사용자에게 저작권 침해가 발생했다고 정보를 제공하게 되며, $RB_i(k)$와 $QB_i(k)$의 기준점블록 내용을 함께 표시함으로써 질의문서 Q의 모든 내용을 읽지 않고도 저작권이 침해된 위치를 정확히 알 수 있게 된다. 침해진단 모듈에서 처리된 정보는 검색엔진모듈로 전달되며 실제로 저작권 문서의 색인 정보 저장과 검색을 처리하게 된다.

III. 구현 및 테스트

본 논문에서 구현된 시스템은 사용자가 인터넷을 통해 저작권 침해 진단 시스템에 접속할 수 있도록 사용자 인터페이스를 제공한다. 사용자는 웹브라우저를 이용하여 웹서비스모듈에 접속할 수 있으며 저작권 문서를 등록할 수 있는 기능 및 저작권 침해 검사를 위한 질의 문서 등록 기능을 구현하였으며, 개발 환경은 표4와 같다.

표 4. 구현 시스템의 개발 환경
Table 4. Environment of the implemented system

항목	사용 기술/환경
CPU	Intel Core I5-4670 CPU 3.40GHz
메모리	4Giga
디스크 용량	500G
OS	Windows Server 2008R2 Standard 64Bit

저작권 침해 문서가 존재하는 경우 질의 콘텐츠 항목을 마우스로 클릭하면 저작권 침해 목록 및 내용을 매칭률과 함께 보여줄 수 있으며 총 문서에 대한 페이지별 문서 분포도를 볼 수 있다.

실제 1,661권의 도서를 대상으로 테스트한 결과 특징수 509,888개, 총 색인시간 826초, 저작권 불법 침해 판단 응답시간 평균 0.345초, 저작권 불법 침해 판단 정확률 100%의 성능을 보였다.

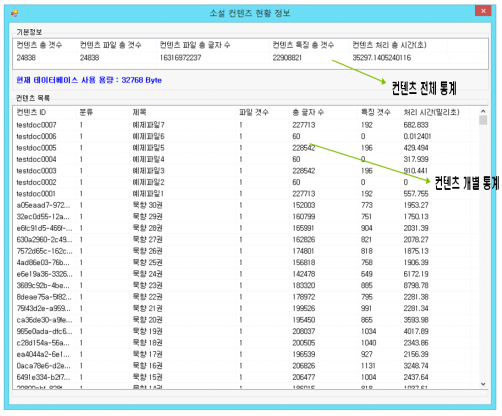

그림 2. SCP 인덱싱 정보
Fig. 2. SCP Information of Indexing

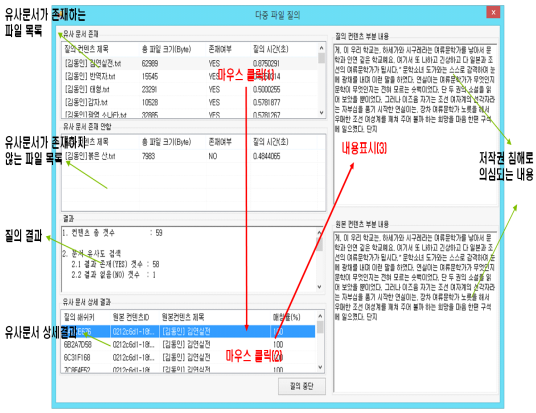

그림 3. 질의결과
Fig. 3. Query Result

V. 결론 및 향후 과제

본 논문은 한국저작권위원회의 저작권기술 개발 사업으로 수행중인 “소셜 저작물의 저작권 보호 및 콘텐츠 메시업 도구 요소기술 개발” 과제의 2차년도 개발결과의 하나로, 윈도우를 이용한 텍스트 기준점 기반의 저작권 침해 판단 시스템을 구현하였다. 본 시스템의 특징은 문장 또는 단락 단위가 아닌, 윈도우 단위의 텍스트 기준점을 이용함으로써 자동으로 기준점을 추출하고, 추출된 기준점을 기반으로 저작권 침해 여부를 판단할 수 있도록 개발 하였다. 또한 텍스트 기준점 정보를 색인하기 위해 검색엔진을 이용함으로써 저작권 침해 판단에서 중요한 진단 속도 향상과 시스템의 확장성을 동시에 제공할 수 있도록 하였다.

이를 통하여 출판자가 창작한 문서 저작물 및 전작책[7]을 등록하기 전에 실시간으로 저작권 침해 여부를 판단할 수 있도록 하여 저작권 침해 예방에 도움이 될 수 있도록 한다.

References

[1] Analysis Report on Copyright Protection, "The type of copyright infringement offence : Comparative analysis of the intellectual property laws" Korea Federation of Copyright Organizations, 2013
[2] "Annual Report on Copyright Protection", Korea Federation of Copyright Organizations, 2013
[3] Myeong-Gi Chae "Copyright issues on the Internet" Korean Institute of Information Scientists and Engineers 21.Vol.30 N0.10.p.21-28
[4] Yong-Jun Hwang "Adaboost-based Gesture Recognition Using Time Interval Window Applied Global and Local Feature Vectors"
[5] Benjamin Hummel, Elmar Juergens, Deniela Steidl, “Index-Based Model Colene Detection” IWSC '11 Proceedings of the 5th International Workshop on Software Clones , pp.21~27, 2011
[6] Sung-Jick Lee, Han-Joon Kim, "Keyword Extraction from news Corpus Using Modified TF-IDF, “Society for E-business Studies" , 2009
[7] Yoon-Ho Kim*, Seong-Hwan Cho** “A Study on the Linkage and integration of UCI (Universal Content Identifier) between ICN (Integrated Copyright Number)” IIBC VOL. 14 No. 5, October 201

※ 본 논문은 문화체육관광부의 저작권기술개발사업에 의거 한국저작권위원회의 정부지원금을 받아 연구되었습니다.
(This research project was supported by Government Fund from Korea Copyright Commission.)

저자 소개

최 경 웅(정회원)

- 1996년 : 호원대학교 정보통신공학과 졸업(학사)
- 1998년 : 전북대학교 영상정보공학과 졸업(석사)
- 2002년 : 전북대학교 정보통신공학과 수료(박사)
- 2004년 ~ 현재 : ㈜아워텍 대표이사

<주관심분야 : 문서 표절, 저작권보호기술, 텍스트마이닝,>

박 순 철(정회원)

- 1982년 : 인하대학교 (석사)
- 1991년 : 루이지아나 대학교 컴퓨터 공학(박사)
- 1993년 ~ 현재 : 전북대학교 컴퓨터 공학부 교수

<주관심분야 : Data Mining & Knowledge Discovery, Artificial Intelligence, Digital Archives and its Application, Topic Dection and Tracking>

양 승 원(정회원)

- 1985년 : 전북대학교 전산통계학과 (학사)
- 1987년 : 전북대학교 전산통계학과 (석사)
- 1995년 : 전북대학교 전산통계학과 (박사)
- 1994년 ~ 현재 : 우석대학교 컴퓨터 공학과 교수

<주관심분야 : 자연어처리, 음성처리, 빅데이터 처리>