인터넷 검색과 형태소분석을 이용한 표절검사시스템의 개발에 관한 연구

황 인 수*

Development of A Plagiarism Detection System Using Web Search and Morpheme Analysis

Insoo Hwang

Abstract

As the World Wide Web (WWW) has become a major channel for information delivery, the data accumulated in the Internet increases at an incredible speed, and it derives the advances of information search technologies. It is the search engine that solves the problem of information overloading and helps people to identify relevant information. However, as search engines become a powerful tool for finding information, the opportunities of plagiarizing have increased significantly in e-Learning.

In this paper, we developed an online plagiarism detection system for detecting plagiarized documents that incorporates the functions of search engines and acts in exactly the same way of plagiarizing. The plagiarism detection system uses morpheme analysis to improve the performance and sentence-based comparison to investigate document comes from multiple sources. As a result of applying this system in e-Learning, the performance of plagiarism detection was improved.

Keywords : e-Learning, Plagiarism, Search Engine, Morpheme Analysis

논문접수일 : 2009년 01월 12일 논문게재확정일 : 2009년 03월 13일

* 전라북도 전주시 완산구 효자동 3가 1200, 전주대학교 미디어정보학부, Tel : 063-220-2757, e-mail : insoo@jj.ac.kr

1. 서 론

인터넷 인프라의 구축이 확대되고 인터넷의 활용이 생활화됨에 따라, 인터넷은 신문, 방송, 잡지 등 대중매체의 기사로부터 개인의 일상을 기술한 일기에 이르기까지 모든 정보의 원천이 되고 있다. 정보량의 증가는 정보의 공유를 통한 새로운 가치의 창출을 가능하게 하였으며, 시간과 공간의 제약을 극복하여 언제 어디서나 원격으로 학습할 수 있는 새로운 교육체제인 e-러닝(Cyber Learning, e-Learning)을 확산시키고 있다. e-러닝은 강의콘텐츠의 개발 및 접근의 용이성과 효율성 등을 바탕으로 자기주도적 학습을 위한 가장 효과적인 수업형태의 하나로 자리잡아가고 있다[임정훈, 1998; 나일주, 1999; 박찬정 등, 2001; 황인수 2008).

그러나 정보의 공유가 긍정적인 효과만을 가져온 것은 아니며, 표절(plagiarism)과 저작권 침해라는 새로운 사회적 문제를 발생시키고 있다. 최근 인터넷에서 가장 많은 정보원천으로 등장하고 있는 블로그의 통계를 작성하는 블로그얌(BlogYam)의 2008년도 4월 블로그 이용실적 통계에 따르면, 본인이 작성한 콘텐츠로만 운영하는 블로그의 비중은 22%에 불과하며, 대부분의 블로거는 신문, 잡지, 연구자료 등을 스크랩하여 운영하고 있는 것으로 나타났다.

교육환경에 있어서, 상당수의 학생들은 과제에 대한 충분한 이해 없이 인터넷에서 수집한 자료를 오려붙이는 짜깁기(Cut and Paste)방법으로 과제를 수행하고 있다. 이에 따라 교수자들은 과제를 손으로 작성하여 제출하도록 요구할 뿐만 아니라, 과제의 복제여부를 파악하기 위해 많은 시간을 소비하는 것을 흔히 볼 수 있다. 보다 심각한 문제는, 인터넷 검색엔진이 연구보고서 등 전문자료와 리포트 등 유료자료를 검색하는 기능을 제공하기 때문에, 인터넷에서 성행하고 있는 리포트 샵으로부터 과제를 구입하여 그대로 제출하는 사례도 점점 더 증가하고 있다는 것이다.

인터넷 자료의 복제는 저작권과 표절의 법적인 문제를 발생시킬 뿐만 아니라 정보화시대에서 정보의 바람직한 활용을 저해하기 때문에, 표절여부를 확인하는 시스템을 개발하는 연구가 수행되어 왔다. 첫째는 연구논문이나 기사 등의 저작권 침해 여부를 확인하기 위해 중앙관리시스템을 구축하는 것이며, 둘째는 표절로 의심되는 문서들 간의 유사도를 분석하는 것이다. 그러나 이러한 방법은 최근에 문제가 되고 있는 인터넷을 표절의 수단으로 삼는 문제에 대해서는 그 해결책을 제공하지 못한다[Liu et al., 2007].

이에 따라 본 연구에서는 과제의 표절여부를 효과적으로 확인할 수 있는 e-러닝시스템을 구축하기 위해 인터넷에서 표절한 자료를 인터넷을 이용하여 검사하는 시스템을 제안한다. 국내에서 가장 대표적인 검색엔진, 뉴스포털, 블로그, 지식, 카페 등을 운영하고 있는 네이버(Naver)와 다음(Daum), 그리고 세계적으로 가장 많은 검색 콘텐츠를 갖고 있는 구글(Google) 등 세 개의 검색엔진을 동시에 검색하여 문서의 표절여부를 검사하는 일종의 인터넷 에이전트이다.

대학에서의 표절 자료의 상당수는 웹문서 보다는 블로그, 카페, 지식, 그리고 유료정보 등을 원천으로 하기 때문에 네이버와 다음이 구글보다 더 좋은 성과를 보일 것으로 예상된다. 최근 다음이 구글과 제휴하여 구글로부터 웹문서 검색결과를 제공받고 있지만, 다음은 구글이 갖지 못한 블로그와 카페 등을 갖고 있으므로 다음과 구글을 모두 검색대상에 포함시키는 것은 타당한 일이다.

본 연구는 메타검색을 이용하여 HWP, DOC, PPT, XLS, TXT, HTM 등 다양한 형식을 갖는 문서의 표절여부를 검사하는 시스템을 개발

할 뿐만 아니라, 표절검사의 성과를 높이기 위해 형태소(Morpheme) 분석을 도입하는 방안을 제안한다. 또한, 표절검사를 위한 문자열 추출 길이에 따른 표절검사 성과의 변화추이 분석, 단일문장 추출과 다중문장 추출의 성과차이 분석, 원문장 분석과 형태소 분석의 성과차이 분석 등 다양한 분석을 통해 표절검사 성과를 향상시키기 위한 방안을 강구한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 표절이론과 이를 검사하기 위해 개발된 시스템에 대해 기술하며, 제 3장에서는 본 연구에서 개발하는 시스템에 대해 기술한다. 제 4장에서는 시스템의 적용사례를 분석하며, 제 5장에서는 본 연구의 성과를 정리한 후 향후 연구방향을 제시한다.

2. 표절에 관한 문헌연구

2.1 표절의 개념

표절의 개념을 파악하기 위해 그 어원을 찾아보면 다음과 같다. 한자에서는 '표독할, 빼앗을 표(剽)'에 '도둑질할 절(竊)'로서 도둑질의 의미를 갖고 있으며, 영어에서는 'Plagiarism'으로서 접두어로 사용한 'plagi-'는 그 어원이 도덕적으로 '부정한' 혹은 '잘못된'이란 뜻으로 부정적인 의미가 내포되어 있다. 국어사전에서는 "다른 사람의 글을 취하여 자기가 쓴 것처럼 행세하는 행위"라고 정의하고 있으며, 저작권심의조정위원회가 펴낸 「저작권표준용어집」에서는 "표절이란 일반적으로 다른 사람의 저작물의 전부나 일부를 그대로 또는 그 형태나 내용에 다소 변경을 가하여 자신의 것으로 제공 또는 제시하는 행위"로 정의하고 있다.

표절은 법률상의 개념은 아니고 정신적 창작물에 대한 절도행위에 대해서 관습적으로 사용되어온 표현이다. 표절은 저작권법으로 보호를 받고 있는 저작물 또는 저작물의 일부를 변경하지 않거나 변경된 형태로 마치 자신이 창작한 것처럼 하여 전용하는 것을 말한다. 이러한 사기행위는 문서위조 또는 저작권침해와 밀접한 관련을 가지며, 일반적으로 저작권법 위반이 된다. 표절은 원래 문학에서 주로 적용되었으나 최근에는 다양한 유형의 저작물 외에 특허, 실용신안, 상표, 의장 등의 산업재산권분야에 이르기까지 그 의미가 확대되고 있다[계승균, 2000].

표절은 원문으로부터 전체 혹은 일부를 취하는 것으로부터 발생하기 때문에, 다음과 같이 두 가지 유형으로 분류할 수 있다. 첫째는 내적(Intra-corpal) 표절로서 동일한 과제나 유사 문서에 대해 내부의 다른 사람의 문서를 복제한 것이며, 둘째는 외적(Extra-corpal) 표절로서 인터넷, 교재, 기사 등 외부의 원천으로부터 문서를 복제한 것이다. 첫 번째 문제에 있어서는 검토 대상이 되는 문서의 개수가 제한적이며 문서간의 유사도에 대한 많은 연구결과가 있으므로 비교적 쉽게 해결될 수 있다. 그러나 두 번째 문제에 있어서는 자료의 원천이 광범위하기 때문에 표절여부를 판정하는 데 많은 어려움이 있다[Liu et al., 2007].

최근의 과제물 작성 방식은 컴퓨터 키보드에 비유하여 [Ctrl-C] [Ctrl-V] 혹은 글을 오려붙인다는 의미로 '짜깁기'로 불린다. 여기서 인터넷은 표절의 중심에 있다. 오동석[2006]은 "관행을 앞세운 침묵의 카르텔이 학계를 지배하고 표절의 기준이 없어 학생들도 인터넷에 기대어 스스럼없이 남의 글을 베낀다."고 주장했다.

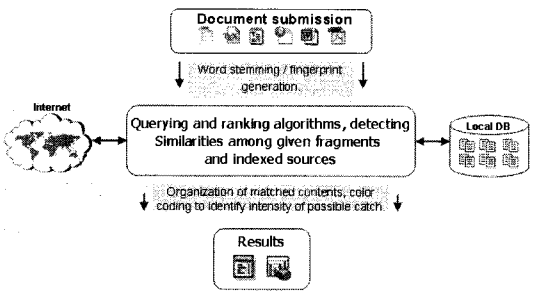
2.2 표절검사시스템

표절검사에서 주된 관심사는 표절에 사용한 교재, 신문, 잡지, 논문, 혹은 인터넷의 원문서를

확인하여 표절여부를 판정하는 것이다. 그러나 표절검사를 위한 통합 데이터베이스 시스템이 구축되어 있지 않기 때문에 문제가 발생했을 때 수작업으로 진행될 수밖에 없는 것이 현실이다.

또한 과제물의 표절검사는 교수가 수작업을 통해 내적 표절을 확인하고 있으나, 그 정확도와 신뢰도는 현저히 낮으며, 외적 표절에 대한 검사는 현실적으로 불가능한 실정이다. 따라서 상당수의 학생들은 인터넷의 리포트 샵에서 수천원에 판매하는 과제물을 그대로 제출하여 성적을 받고 있다.

이에 따라, 턴잇인(TurnItIn), Eve2, KURE-POLS 등 인터넷 정보와 자체 구축한 데이터베이스를 이용하여 표절여부를 검사하는 시스템을 구축하고 있는데, 이의 동작원리를 그림으로 나타내면 <그림 1>과 같다.



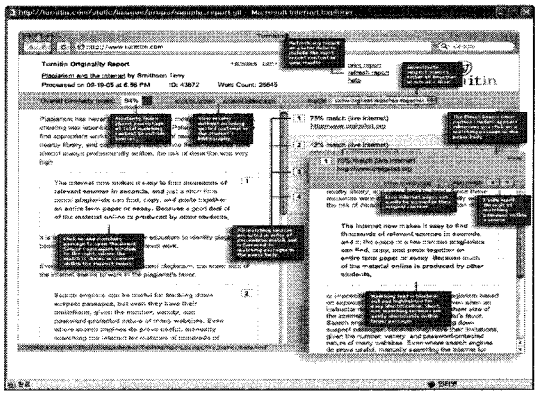
〈그림 1〉 표절검사의 개념도[Manber, 1994]

표절검사시스템에서 사용하는 알고리즘은 크게 속성계수(Attribute Counting) 방법과 구조분석(Structure Metric) 방법으로 구분된다[조동욱 외, 2003]. 여기서 속성계수법은 문서에 사용된 단어의 출현빈도와 유사성을 검사하는 방법이며, 구조분석은 단어들의 정확한 일치가 아닌 토큰 스트링(Token String)의 유사성을 계산하는 방법이다. 따라서 일반 문서의 표절여부를 검사할 때에는 속성계수법이 많이 사용되고, C나 JAVA 등 프로그래밍 언어로 작성된 프로그램의 표절여부를 검사할 때에는 구조분석방법이 많이 사용된다.

(1) 턴잇인(TurnItIn)

미국 아이패러다임(Iparadigms. LLC)사는 MS 워드나 TXT, HTML 등 다양한 파일형태로 작성된 리포트나 논문을 전 세계 학술관련 데이터베이스(DB)와 비교하여 표절여부를 퍼센트(%)로 제시해 주는 턴잇인 시스템을 개발하였다. 120억개 이상의 웹문서와 4,000만건 이상의 학생 리포트, 1만여 건 이상의 학술저널과 뉴스정보를 기반으로 표절여부를 검색해 주는 시스템으로, 세계적으로는 9,000여개 대학과 연구기관 등이 이 프로그램을 도입해 사용하고 있다.

국내에서도 ICU와 KAIST 등이 이 시스템을 사용하고 있다. 그러나 이 시스템은 검사할 파일을 업로드하기 때문에 표절여부를 확인하기까지 수 분에서 수십 분의 시간이 소요될 뿐만 아니라 사용료를 지불해야 한다. 뿐만 아니라, 한글 서비스가 이루어지지 않는다는 점은 국내에서 적용할 때 가장 큰 한계점이 될 것이다.

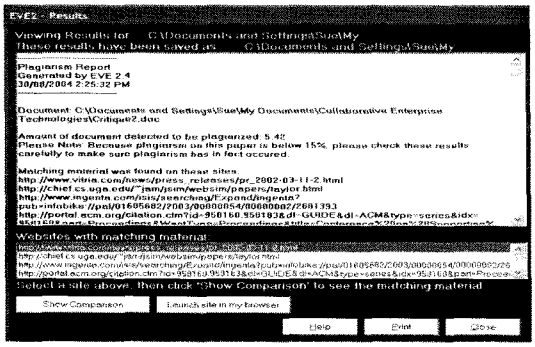


〈그림 2〉 턴잇인(TurnItIn) 검사보고서의 예

(2) Eve2

Eve2(Essay Verification Engine)는 인터넷 검색을 통해 표절여부를 검사하는 시스템으로 턴잇

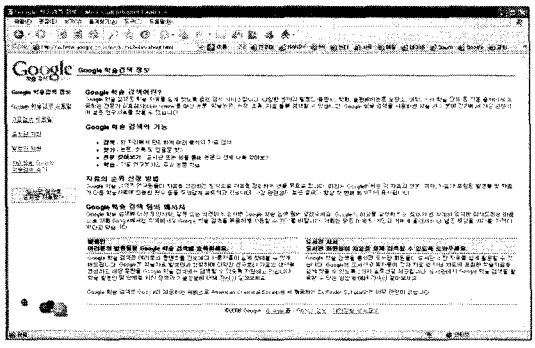
인이 웹 기반 프로그램으로 구현된 반면에 Eve2는 윈도우즈용 프로그램으로 개발되어 PC에서 사용한다. 텃잇인에 비해 빠르고 저렴하지만, 하나 이상의 문서로부터 표절을 한 경우 표절확인이 용이하지 않으며 텃잇인과 같이 다양한 형태의 검사결과보고서를 제공하지 못하는 단점이 있다.



〈그림 3〉 Eve2 표절검사 프로그램의 실행결과

(3) 구글 학술검색

구글(Google)은 정보검색분야의 강점을 살려서 학술정보만을 검색하는 '학술검색 β'를 개발하여 시험 운영하고 있다. 구글 학술검색은 학술 자료를 쉽게 찾도록 돕는 검색 서비스로서, 다양한 분야의 발행인/출판사, 학회, 출판예비논문 보관소, 대학, 기타 학술 단체 등 각종 출처에서 제공하는 전문가 상호심사(Peer Review)를 마친 논문, 학술논문, 서적, 초록, 자료 등을 검색할 수 있다.

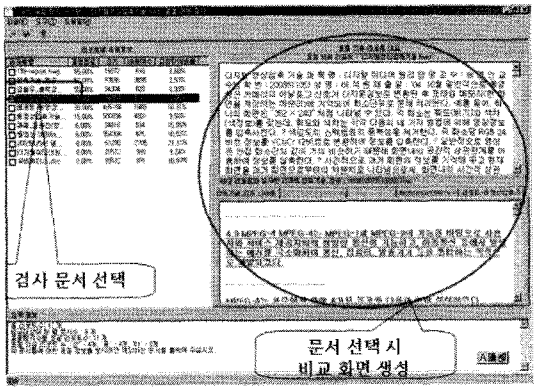


〈그림 4〉 구글 학술검사의 예

그러나 이 시스템은 구축 및 시범운영단계에 있으며 학술정보의 검색을 목적으로 하기 때문에 과제물 등의 표절을 검사하는 시스템으로 활용하기에는 적합하지 못하다.

(4) KUREPOLS

KUREPOLS(Korea University Report Police System)는 고려대학교 정보통신대학에서 개발한 표절검사시스템으로 2006학년도 1학기부터 활용하고 있으며, 관련 내용에 대한 특허를 출원하였다[임해창 외 2, 2006]. 이 시스템은 표절검사 대상 문서에서 추출한 복수개의 검색어로 인터넷을 검색하여 관련 웹문서를 추출한 후, 검사문서와 웹문서에서 추출한 색인어를 이용하여 유사도를 계산함으로써 표절여부를 판정한다.



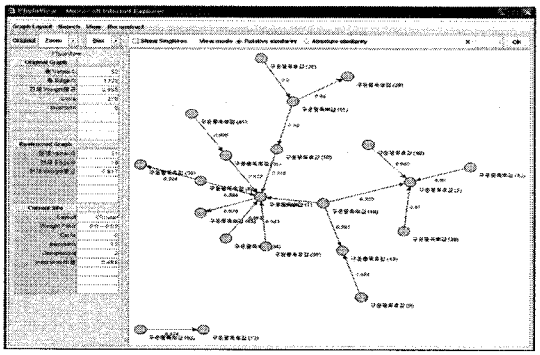
〈그림 5〉 KUREPOLS의 내적 표절검사의 예

검색어 및 색인어를 추출할 때 형태소를 이용함으로써 표절여부 판정의 정확도를 높이고 있으나, 검사대상 문서가 여러 개의 웹문서를 이용하여 작성되었거나 혹은 전문을 검색할 수 없는 유료정보를 이용한 경우에는 검사가 용이하지 않다는 단점이 있다.

(5) DEVAC

DEVAC(Document EVolution Analyzing Center)은 생물의 유전자를 검사하는 방식을 응용하여 어절단위로 각종 문서의 유사도를 측정하기 때문에 매우 빠른 속도로 검사를 수행한다는 장점이 있다[류창건 외, 2007]. 단행본 2권의 표절 여부를 확인하는 데 약 1.5초, A4용지 5장짜리 리포트 100명분을 확인하는 데 약 5분이 소요된다. 최근에는 대하소설 '태백산맥'을 표절해 물의를 빚고 있는 유명작가의 소설을 분석하여 총 9곳의 유사 대목을 확인해 내기도 했다.

윈도우즈용 애플리케이션으로 제공되는 DEVAS의 Phylogenetic Tree View 모듈은 표절 탐색의 결과를 이용하여 문서들 사이의 표절 흐름을 추적할 수 있다. 어떤 문서를 통해 표절이 진행되어 왔는지를 생물정보학의 계통 발생론적 원리를 이용하여 그래프로 쉽게 알아 볼 수 있는 기능을 제공한다.



〈그림 6〉 문서간의 표절관계를 보여주는 그래프

3. 표절검사 시스템 개발

3.1 문장단위 분석

기존의 연구들은 검사대상 문서와 비교대상 문서에서 추출한 색인어가 유사한 정도에 따라 표절여부를 판정하는 것으로서, 일반적으로 하나의 문서를 표절하며 비교대상 문서가 제한적인 내적 표절문제에서는 쉽게 적용될 수 있다.

그러나 여러 개의 문서를 짜깁기하여 문서를 작성한 경우에는 표절여부를 판단하는 것이 어려울 뿐만 아니라, 문서에 포함된 전체 문장에서 색인어를 추출하기까지 많은 시간이 소요된다. 특히 외적 표절문제에서는 비교대상이 되는 모든 문서를 색인화 하여 검사하는 것은 물리적으로 불가능한 일이다.

본 연구는 상대적으로 표절여부의 확인이 어려운 외적 표절을 검사하는 시스템의 개발에 초점을 맞추고 있다. 따라서 본 연구에서는 문서단위의 비교가 아닌 문장단위 비교 방법을 채택하였으며, 비교대상이 되는 문서의 전체 문장을 분석하지 않고 인터넷 정보검색을 통해 제시되는 3~4줄의 검색결과만을 분석한다. 이 방법이 갖는 장점은 검색엔진의 검색결과에 따라 다시 해당 웹문서를 방문하는 추가적인 검색이 불필요할 뿐만 아니라, 문장단위로 서로 다른 여러 개의 문서를 표절한 경우에도 표절여부를 쉽게 검사할 수 있다는 것이다.

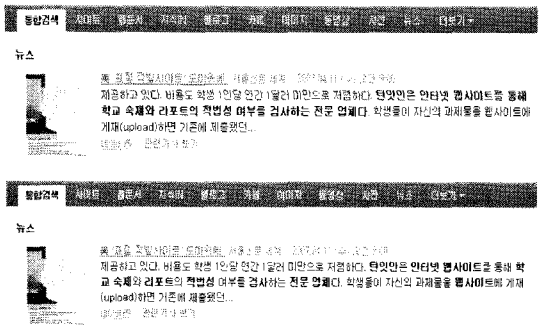
문장단위 검색에 있어서 문장을 어떻게 구분할 것인가는 매우 중요한 문제로서, 문장은 일반적으로 마침표(.), 느낌표(!), 물음표(?) 등의 문장부호에 의해 구분된다. 그러나 문장부호를 2개 이상 연속으로 사용하거나 날짜 표기에서 마침표(.)를 삽입하는 등 문장구분으로 적합하지 않은 예외사항이 발생하기 때문에 문장을 구분할 때 주의해야 한다. 또한, 문장이 너무 짧은 경우에는 표절검사를 위한 질의어로 의미가 없기 때문에 문장 내에 형태소의 개수가 작은 문장은 표절검사에서 제외한다.

3.2 원문장 질의

현재 운영되고 있는 대부분의 검색엔진은 키

워드 검색 뿐만 아니라 자연어 검색을 지원한다. 이것은 사용자가 일상에서 사용하는 자연어의 형태로 검색엔진에 질의하면 검색엔진은 질의어로부터 형태소를 추출하여 데이터베이스를 검색하는 것이다. 즉, 사용자는 자연어로 질의를 하지만 검색 엔진은 종전과 동일한 키워드 방식으로 질의를 처리한 후, 자연어로 처리한 것처럼 결과를 제시하는 것이다.

예를 들어 "턴잇인은 인터넷 웹 사이트를 통해 학교 숙제와 리포트의 적법성 여부를 검사하는 전문 업체다"는 자연어 문장을 네이버에서 검색하면 <그림 7>과 같다. 이 문장에서 색인어가 될 수 있는 단어들을 추출하고 이들의 순서를 변경하여 "웹 사이트 턴잇인 인터넷 숙제 학교 적법성 리포트 검사 업체 전문"과 같은 다중 단어검색을 실시하더라도 동일한 결과를 도출하게 된다.



〈그림 7〉 네이버의 자연어 및 키워드 검색 결과

이것은 표절검사 프로그램에서 사용하는 형태소 추출방식이 검색엔진에서 사용하는 것과 동일한 방식일 경우에는 표절검사를 위해 형태소를 추출하는 것이 의미가 없음을 보여준다. 만일, 서로 다른 형태소 추출방식을 사용하면서 형태소 변환 문장을 질의어로 사용하게 되면, 주요 단어의 누락이나 불일치가 발생하여 오히려 바람직하지 못한 검색결과를 도출할 수 있다.

따라서 본 연구에서는 검색엔진의 질의어로 검사대상 문서에 포함된 원문장을 그대로 사용한다. 다만, 검색된 결과로부터 표절여부를 검증하기 위해서는 형태소를 추출해야 하는데, 본 연구에서는 강승식[1996]이 개발한 HAM을 이용하였다.

3.3 정보추출규칙

본 연구는 원문장과 인터넷 정보검색결과를 비교하여 표절여부를 판정하기 때문에 검색엔진의 검색결과인 웹 문서로부터 특정 부분의 텍스트 정보만을 추출해야 한다. 이와 같이 정보원으로부터 원하는 정보를 추출하기 위한 규칙이나 프로그램을 정보추출규칙(Wrapper)이라고 하며[박상위 2002], 웹 문서마다 그 구성 방식이나 내용이 다르기 때문에 추출하고자 하는 정보의 종류나 웹문서의 구성에 따라 서로 다른 정보추출규칙을 적용해야 한다[Kushmerick et al., 1997].

이에 따라 본 연구에서 검색대상으로 하는 네이버, 다음, 그리고 구글 검색엔진을 위해 <표 1>과 같은 정보추출규칙을 설정하였다. 이들 규칙은 수시로 변경될 수 있기 때문에 인터넷 서버에 생성한 후, 클라이언트 프로그램에서 검색시마다 참조하도록 구성하였다.

〈표 1〉 검색엔진별 정보추출 규칙

엔 진	시작태그	종료태그
네이버	<dd>	</dd>
다 음		
구 글		

3.4 표절지수

기존에 개발된 대부분의 표절검사 시스템은

표절여부를 판정하기 위해 검사대상 문서와 비교대상 문서간의 유사도를 사용한다. 문서간의 유사도를 평가하는 대표적인 모형인 벡터공간 모형(Vector Space Model)은 문서의 키워드를 다차원 벡터공간에 매핑하여 문서별로 벡터를 만들어 둔 후, 각 벡터의 사이각을 비교하는 방식으로, 코싸인(Cosine) 함수를 사용하여 사이각이 작을수록 높은 연관도를 갖는 것으로 평가한다. 그러나 이 모형은 긴 문서들은 유사도 값이 작기 때문에 제대로 표현되지 않으며, 벡터공간 표현에서는 단어들이 나타나는 순서가 무시되는 등의 단점이 있다. 특히, 문서와 문서를 비교하기 때문에 본 연구에서와 같이 하나의 문서를 여러 개와 비교하는 것은 불가능하다.

이에 따라 본 연구에서는 임의로 선택된 문장에 대한 반복적인 정보검색 및 분석의 결과로부터 표절지수(CPI, Cut and Paste Index)를 산정하는 새로운 방법을 제안한다. CPI의 계산과정을 의사코드로 나타내면 <그림 8>과 같다.

<그림 8>에서 질의의 Qi와 문서 Dij간의 유사도는 Qi의 단어 Wik가 문서 Dij에 포함되어 있는 개수의 비율과 출현순서가 불일치하는 단어의 개수로부터 계산된다. 여기서 가중치 α는 임의로 설정할 수 있으나 본 연구에서는 0.5로 결정하였다. 또한, 표절지수(CPI)를 계산할 때 유사도 Si에 유사도 한계치 T를 적용하는 것은 문서간 비교의 경우에는 일치하는 색인어의 개수가 모두 의미를 갖지만, 본 연구에서와 같이 검색엔진에 질의한 결과에서는 다른 내용의 문장인 경우에도 일치하는 단어들이 포함될 수 있기 때문이다. 여기서 T의 값은 임의로 설정할 수 있으나, 본 연구에서는 시뮬레이션을 통해 0.5로 설정하였다.

3.5 표절검사 과정

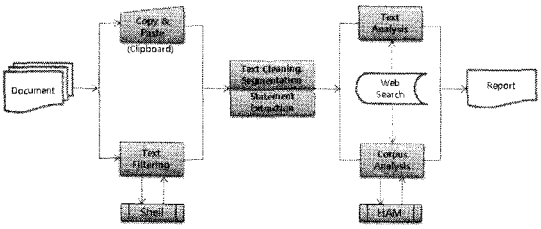
인터넷을 이용한 외적 표절은 검색엔진에 몇 개의 키워드를 입력하여 검색된 문서들을 짜깁기하는 방식으로 이루어진다. 따라서 표절을 검사하는 시스템은 표절과정을 역으로 이용하여 표절이 예상되는 문서로부터 일부의 문장을 취하여 검색함으로써 원문서를 확인한다. 본 연구에서는 외적 표절검사를 위한 대부분의 시스템과 마찬가지로 검색엔진을 이용하지만, 구체적

1. 다음을 N회 반복하여 수행한다.
   1.1 검사대상 문서내 임의의 문장으로부터 길이 M의 검색문자열 Qi를 추출한다.
   1.2 검색엔진에서 Qi로 검색한 결과를 문자열 Ri에 할당한다.
   1.3 검색결과 Ri로부터 차례대로 웹문서 Dij의 텍스트를 추출하여 다음을 반복한다.
      1.3.1 Qi에서 차례대로 단어 Wik를 추출한다(토큰화 또는 형태소 추출).
      1.3.2 Dij에 단어 Wik가 포함되어 있는지 검사한다.
      1.3.3 Qi의 각 단어 Wik가 웹문서 Dij에 출현하는 비율과 배열의 순서로부터 Qi와 Dij의 문서간 유사도 Sij를 계산한다.
         유사도 = (출현단어의 개수-α · 순서불일치 단어의 개수)/검사단어의 개수
   1.4 Ri에 속한 웹문서 Dij에서 가장 큰 유사도를 Qi의 유사도 Si로 설정한다.
2. Si가 유사도 한계치 T보다 큰 경우만 값을 갖도록 다음과 같이 Si를 다시 계산한다.
   Si = Max(0, (Si-T)/(1-T))
3. Si의 평균 값으로부터 CPI를 계산한다.

〈그림 8〉 표절지수를 계산하는 과정

인 웹페이지를 검색하지 않고 검색엔진의 검색 결과를 단위문장단위로 비교하여 표절여부를 판정한다.

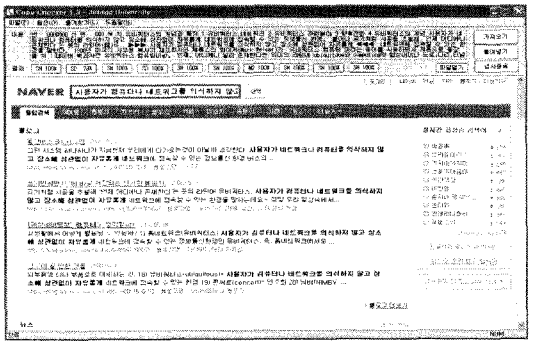
<그림 9>는 표절 검사가 이루어지는 과정을 보여주고 있는데 이를 개략적으로 설명하면 다음과 같다. 먼저, 표절 검사대상 문서를 클립보드에 복사하거나 혹은 파일을 직접 불러오는 방식으로 검사할 문서의 텍스트를 읽어온 후, 검사를 수행하기 위해 원본 텍스트를 정리하고 문장을 추출하는 과정을 수행한다. 다음으로 인터넷 검색을 수행한 결과에 대해 원문장 분석 또는 형태소 분석을 통해 CPI를 계산하고 검사결과를 제시한다.



〈그림 9〉 표절검사의 흐름도

<그림 10>은 본 연구에서 개발한 표절 검사 시스템의 수행결과를 보여주고 있다. 시스템은 Visual Studio 2008에서 Visual C++와 C# 프로그래밍 언어를 이용하여 개발하였으며, 프로그램 상에 웹문서를 그대로 보여줄 수 있도록 Visual C++의 CHtmlView를 기반 클래스로 작성하였다. 검사할 문장을 복사한 후 [붙여넣기] 버튼을 클릭하거나, 혹은 [가져오기] 버튼을 클릭해서 파일을 읽어오는 것만으로 모든 검사가 수행되므로 매우 편리하게 사용할 수 있다.

이 시스템은 검사대상 문서에서 임의로 10개의 문장 및 문자열을 추출하여 네이버, 다음, 구글 검색엔진을 검색한 결과에 대해 원문장 분석 및 형태소 분석을 실시하여 표절지수(CPI)를 계산한다. 이 때, 사용자는 검색에 적용할 검색엔진, 추출할 문자열의 길이, 추출할 문장에 포함될 최소한의 형태소 개수, 단일문장 혹은 다중문장 추출여부, 인용부호내 문장의 추출여부, CPI를 계산하기 위한 유사도 한계값 등을 설정할 수 있다. 표절 검사가 완료된 후, [파일열기] 버튼을 클릭하면 표절 검사를 실시한 문서가 화면에 나타나며, [결과] 버튼을 클릭하면 검색엔진의 검색결과를 직접 확인할 수 있기 때문에 표절여부를 쉽게 판정할 수 있다.



〈그림 10〉 문서의 표절검사 수행결과

4. 적용 및 사례 분석

4.1 사례개요

본 연구에서 개발한 표절검사 시스템의 성과를 평가하기 위해 e-러닝 강좌를 수강하는 178명의 학생들이 제출한 145개의 과제를 사용하였다. 이 강좌는 경기, 강원, 대전, 전북, 전남, 대구, 울산, 경남 등 8개 지역의 대학으로 구성된 RUCK 컨소시엄에서 100% 원격으로 진행하는 공동운영 강좌이다. 따라서 학생들 간의 물리적인 교류가 전혀 없기 때문에 내적 표절문제는 거의 없지만, 강의가 인터넷을 기반으로 이루어지므로 인터넷 문서를 이용한 외적 표절문제는 매우 심각한 상태이다.

제목 : 인터넷 실명제에 대한 기사 작성
내용 : 원격으로 진행한 토론내용과 인터넷 등의 자료를 참조해서 인터넷 실명제에 대한 기사를 A4 용지 1페이지로 작성하여 제출하기 바랍니다. 보고서의 형태라기보다는 사실을 보도하는 형식으로서, 인터넷 실명제의 개념을 설명하기보다는 흥미를 유발할 수 있는 최신의 사례나 뉴스꺼리를 중심으로 기술하기 바랍니다.
주의 : 인터넷으로부터 단 한 문장이라도 동일한 문장으로 COPY를 한 경우에는 과제점수를 0점 처리합니다.

〈그림 11〉 '인터넷 실명제'에 대한 과제부과 내용

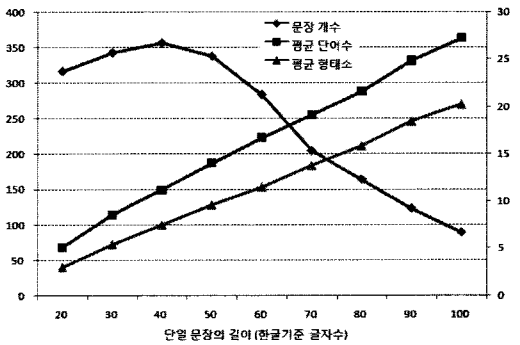
본 연구에서 분석의 대상으로한 과제는 '인터넷 실명제'에 대한 신문/잡지 기사를 작성하는 것이다. 이 주제는 이미 원격수업에서 강의콘텐츠로 제공되었을 뿐만 아니라 학생들 간에 충분한 토론이 이루어졌기 때문에 학생들은 과제를 위한 충분한 지식을 갖고 있다. 그러나 <표 2>에서 보는 바와 같이, 인터넷 정보검색에서는 이 주제와 관련하여 방대한 양의 웹문서, 지식정보, 그리고 과제 파일 등을 제공하기 때문에 학생들의 표절 욕구는 강할 수밖에 없다.

〈표 2〉 '인터넷 실명제' 검색 통계(2008. 6. 7)

분 류	NAVER	DAUM	GOOGLE	합 계
웹문서	82,863	210,000	1,120,000	1,412,863
블로그	7,712	5,096	273	13,081
카 페	2,582	9,416		11,998
뉴 스	2,280	2,982	411	5,673
게시판	1,893	2,049		3,942
지 식	2,981	289		3,270
전문자료	303	413	101	817
이미지	135	99	510	744
동영상	118	6		124
합 계	100,867	230,350	1,121,295	1,452,512

제출된 과제로부터 텍스트를 추출하여 분석한 결과, 과제당 평균 문장의 수는 17개, 문장당 평균 단어의 개수는 13개로 나타났다. <그림 12>는 문장의 길이에 따른 출현빈도, 평균 단어 및 형태소의 개수를 보여주고 있다. 여기서, 형태소는 언어의 형태론적 수준에서의 최소단위로서, 본 연구에서는 명사형태를 갖는 자립형태소를 분석의 대상으로 하였다.

약 30~50글자로 구성된 문장이 가장 많았으며 60글자 이상이 될수록 출현빈도는 현저히 감소하는 것을 볼 수 있다. 문장에 포함된 단어의 개수와 형태소의 개수는 문장의 길이에 비례하여 증가하는데 문장내 단어에서 형태소가 추출되는 비율은 약 75%로 나타났다.



〈그림 12〉 문장의 길이에 따른 빈도, 단어 및 형태소의 개수

4.2 결과분석

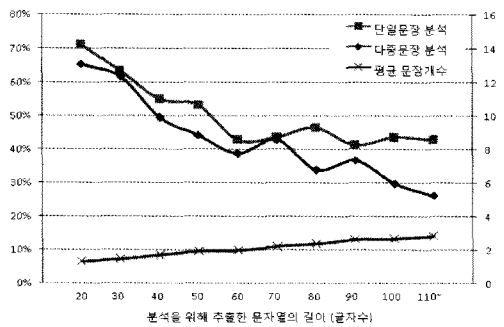
(1) 단일문장 분석

본 연구에서는 표절검사 대상문서가 여러 개의 문서를 표절한 경우에도 표절여부를 확인할 수 있도록 단일문장에서 표절검사 문자열을 추출한다. 이것은 표절검사 대상 문서가 여러 개의 문서를 짜깁기 한 경우 다중문장을 포함하는 문자열로 검사를 하면 표절판정을 받을 확률이 매우 낮기 때문이다.

따라서 본 연구에서는 표절검사 문자열을 단일문장에서 추출한 경우와 다중(혼합) 문장에서 추출한 경우의 유사도를 분석하였다. 학생들이

제출한 145개의 과제에 대해서 단일문장과 다중문장을 포함하는 문자열을 각각 10개씩 임의로 추출하여 유사도를 평가하였다. 여기서 추출할 문자열의 길이는 20글자로부터 110글자까지 변화시키면서 임의로 추출하였으며, 유사도는 추출한 문자열에 속한 단어가 웹문서에 포함되어 있는 비율로 계산하였다.

<그림 13>은 표절검사를 위해 추출한 문장의 길이가 변화함에 따라 단일문장 분석과 다중문장 분석에 따른 유사도의 변화를 보여주고 있다. 여기서, 평균문장의 개수는 검사를 위해 추출한 문자열의 길이에 따라 평균적으로 몇 개의 문자가 포함되는지를 그림으로 나타낸 것이다. 문장의 길이가 길어질수록 유사도가 감소하는 것을 볼 수 있는데, 이것은 전체문장을 표절하기보다는 문장의 일부를 표절하는 현상으로 분석되거나 혹은 검사 문자열이 짧을 경우 표절하지 않은 문장이 우연히 동일한 구성을 갖고 있어서 표절로 판단되는 현상이 있을 수 있다.



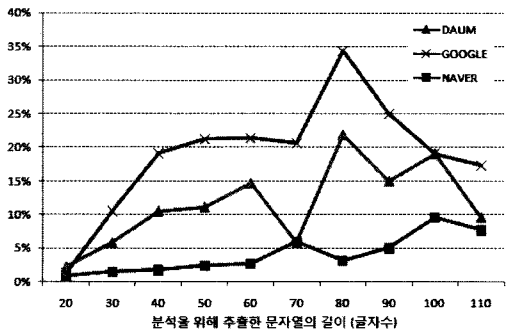
〈그림 13〉 단일문장 추출과 다중문장 추출의 표절검사 성과

이 분석의 결과는 두 가지의 의미 있는 결과를 제공한다. 첫째는 그림에서 보는 바와 같이, 모든 영역에서 단일문장 분석의 성과가 다중문장 분석의 성과보다 높게 나타나고 있다. 이에 따라, 단일문장과 다중문장 추출결과 각각 1,450개의 데이터에 대해 SAS 통계프로그램을 이용하여 T 검정을 실시한 결과, $p < 0.001$에서 유의한 차이가 있는 것으로 나타났다. 따라서 다중문장으로 구성되는 문서간의 유사도를 평가하는 전통적인 방법보다 문서내의 단일문장의 유사도를 평가하는 것이 보다 높은 표절검사 성과를 도출할 수 있음을 알 수 있다. 또한 이러한 연구결과는 학생들이 과제를 작성할 때 2개 이상의 문서를 짜깁기 현상이 있음을 보여주는 증거가 된다.

둘째는 단일문장 분석의 경우 검사를 위해 추출한 문자열의 길이가 일정한 길이를 넘으면 문장의 유사도가 일정한 수준을 유지한다는 것이다. 이것은 다른 문서의 문장을 표절할 때 문장의 일부가 아닌 문장 전체를 표절하는 현상을 보여주는 것으로서, 본 연구에서는 약 60글자 이상일 때 표절검사의 성과가 안정화되는 것으로 나타났다. 따라서 단일문장을 표절검사에 사용하지만 문장의 길이가 60글자 이상인 경우에는 문장으로부터 임의로 60글자를 추출하여 검사를 수행한다.

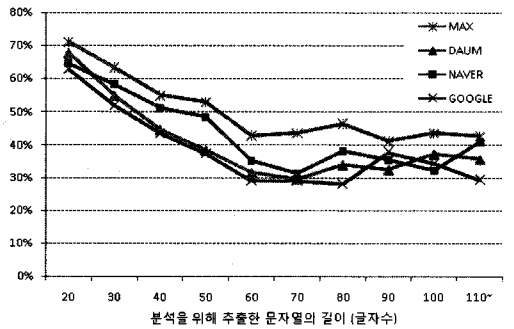
표절검사 문자열의 길이를 결정할 때 고려해야 할 부가적인 사항으로는 검색엔진이 질의에 대한 검색결과를 제시하지 못하는 검색실패율이다. 구글은 32단어 까지만 검색이 가능하며, 다음도 문자열이 길어지면 검색을 실패하는 현상이 두드러지게 나타난다. <그림 14>는 문자열의 길이에 따라서 각 검색엔진이 검색결과를 제시하지 못하는 비율의 추이를 보여주고 있다. 여기서 검색성공률이 100%가 되지 못하는 것은 문서 자체에 오타나 새로운 용어가 포함되어서 검색엔진이 관련정보를 찾지 못하는 경우가 발생하기 때문이다.

그림에서 보는 바와 같이, 구글과 다음은 검사문자열의 길이가 증가할수록 검색실패율이 현저히 증가하며 60글자를 초과하면 검색실패율이 불안정한 것으로 나타났다.



〈그림 14〉 문자열 길이에 따른 검색엔진별 검색실패율(%)

다음으로, 표절검사를 위해 추출한 문자열의 길이에 따라 각 검색엔진의 표절검사 성과를 그림으로 나타내면 <그림 15>와 같다. 분석 결과, 국내에서 네이버가 검색시장의 약 80%를 점유하고 있음에도 불구하고, 표절검사에서는 다음과 구글도 유사한 검사결과를 도출하였다.

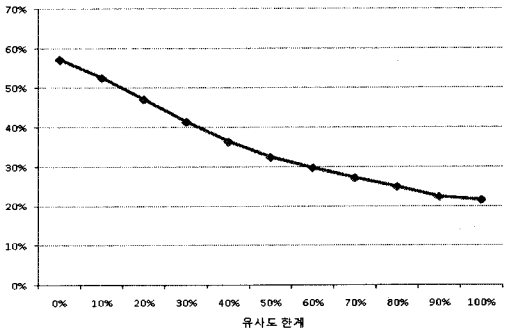


〈그림 15〉 검색엔진별 표절검사 결과

(2) 유사도 한계

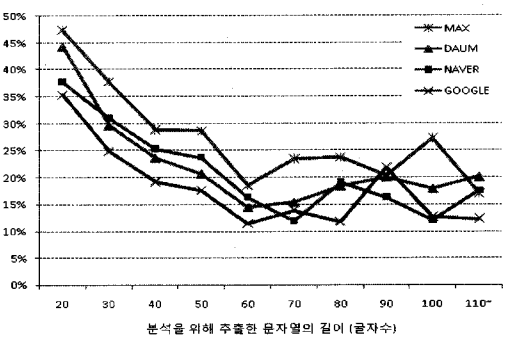
본 연구에서는 기존의 연구와는 달리 문서간의 유사도를 계산하지 않고 표절검사를 위해 추출한 문자열과 검색엔진의 검색결과 문자열을 비교하여 일치하는 단어의 비율에 따라 유사도를 계산한다. 따라서 문장을 전혀 표절을 하지 않은 경우에도 몇 개의 단어는 일치할 수 있기 때문에 유사도 한계 또는 임계치(Threshold)의 개념을 도입할 필요가 있다. 여기서 유사도가 임계치보다 작으면 0, 임계치보다 크면 0~100%로 재계산된다.

$$\text{유사도}' = Max\left(0, \frac{\text{유사도} - \text{임계치}}{100 - \text{임계치}}\right)$$



〈그림 16〉 유사도 한계(임계치)에 따른 유사도 ′의 변화

본 연구에서는 임계치의 변화에 따라 유사도′가 변화되는 추이를 분석하기 위해 컴퓨터 시뮬레이션을 실시하였으며, 그 결과를 그림으로 나타내면 <그림 16>과 같다. 임계치가 커질수록 유사도′가 거의 비례하여 감소하고 있기 때문에 이론적으로 임계치를 설정하기는 어려울 것으로 보이지만, 본 연구에서는 기울기의 변화가 나타나는 50%를 임계치로 설정하였다.

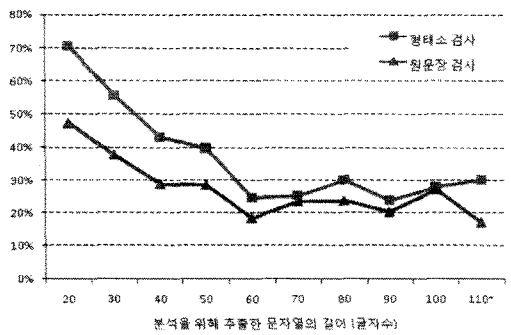


〈그림 17〉 임계치가 0.5일 때 표절지수(CPI)의 변화추이

<그림 17>은 임계치를 50%로 할 때 문자열의 길이에 따라 표절지수(CPI)가 어떻게 변화되는지를 보여준다. 표절지수의 변화패턴은 <그림 15>에서 보여주고 있는 유사도의 변화와 거의 동일한 모습을 보이고 있으며, 검사를 위해 추출하는 문자열의 길이가 60글자에서 표절지수(CPI)의 최대값은 약 18%를 나타내고 있다. 이것은 검사대상 과제물에 포함된 문장의 18%가 표절되었음을 의미한다.

(3) 형태소 분석

문장내 단어의 순서를 일부 변경하거나 문체를 변경하더라도 동일한 문장구조와 표현을 갖고 있으면 표절로 인정되기 때문에 원문장에 나타난 단어를 그대로 비교하는 것보다 형태소를 추출하여 비교하는 것이 표절검사의 성과를 높일 수 있다. 따라서 본 연구에서도 표절검사를 위한 문자열에서 형태소를 추출하여 검색엔진의 검색결과와 비교하는 형태소 분석을 실시하였으며 그 결과를 그림으로 나타내면 <그림 18>과 같다. 여기서 형태소 추출은 강승식[1996]이 개발한 HAM(Hangul Analysis Module)을 활용하였다.



〈그림 18〉 원문장 검사와 형태소 검사의 표절지수 비교

문자열의 길이변화에 따른 형태소 검사결과는 원문장 검사와 거의 동일한 패턴을 보이고 있으며, 문자열의 길이가 60글자 이상인 경우 큰 변화가 없는 것으로 나타났다. 또한 문서에 대한 표절 판정율은 원문장 검사보다 전반적으로 높게 나타나고 있으며, 문자열의 길이 60에서는 약 7%가 상승하여 25%의 문서가 표절된 것으로 판정하였다. 따라서 형태소 검사는 보다 정확하고 엄격한 표절검사를 수행하기 때문에 본 시스템은 형태소 검사의 적용여부를 교수자가 결정할 수 있도록 하였다.

5. 결 론

인터넷 정보량의 증가는 정보검색기술의 발전을 가져왔으며, 인터넷은 모든 정보의 원천이 되고 있다. 이것은 정보의 공유를 통한 새로운 가치의 창출을 가져와 인류의 지식축적에 크게 기여하고 있다. 그러나 정보량이 증가하고 정보검색기술이 발달할수록 기술을 부적합한 방법으로 사용하는 사람들이 증가하고 있어 인터넷을 매체로 하는 표절이 방송/신문/잡지 기사, 학술논문, 교재, 그리고 학생들의 과제물에 이르기까지 사회적 문제로 대두되고 있다.

과제물의 경우 웹문서, 블로그, 카페 등의 정보를 짜깁기하여 제출하는 사례가 급증하고 있으며, 최근에는 과제물을 판매하는 리포트 샵이 성황을 이루고 있다. 이러한 표절은 그룹 내에서 표절이 이루어지는 내적 표절보다는 외부의 다양한 매체를 표절의 도구로 활용하는 외적 표절의 형태로 나타나기 때문에 표절여부를 검사하는 것은 매우 힘든 일이다. 인터넷에서 구입한 과제물로 A+의 학점을 취득하는 것은 불합리한 일이므로, 표절을 검사하는 시스템에 대한 연구가 점점 더 많아지고 있다. 그러나 기존 연구의 대부분은 내적 표절문제를 해결하기 위해 문서간의 유사도를 계산하는 데 초점이 맞추어져 있으며, 외적 표절문제를 다루기 위해 인터

넷 검색엔진을 활용하는 경우에도 특정 문서를 검색하여 그 유사도를 계산하는 방식을 취하고 있다. 이러한 방법은 여러 개의 문서를 표절한 경우에는 표절여부를 확인하는 것이 거의 불가능하게 된다.

따라서 본 연구에서는 문서단위가 아닌 문장단위의 검사를 통해 표절을 확인하는 방법을 채택했다. 사례에 대한 분석에서 표절검사 문자열을 단일문장에서 추출한 경우가 다중문장에서 추출한 경우보다 훨씬 더 높은 표절검사 성과를 나타낸 것은 학생들이 여러 개의 문서를 표절하고 있음을 보여주는 것이다. 또한, 표절검사결과를 평가하기 위한 지표로 CPI(Cut and Paste Index)를 제안하였으며, 표절검사를 위해 추출하는 문자열의 길이에 따른 표절검사의 성과를 컴퓨터를 이용한 다양한 시뮬레이션을 통해 확인하였고, 형태소 검사가 표절검사의 성과를 향상시킬 수 있음도 보였다.

본 연구에서 개발한 표절검사 시스템은 문서를 클립보드에 복사한 후 [붙여넣기] 버튼을 클릭하거나, 혹은 [가져오기] 버튼을 클릭하여 파일을 선택하는 것만으로 표절검사가 완료되며, 표절검사의 결과를 CPI로 제시할 뿐만 아니라 검색엔진의 검색결과를 평가자가 직접 확인할 수 있다는 장점이 있다. 향후의 연구에서는 여러 개의 문서를 입력받아 동시에 표절검사를 수행하며 내적 표절여부도 함께 검사하는 방안에 대해 연구할 예정이다.

참 고 문 헌

[1] 강승식, 장병탁, "음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기", *정보과학회 논문지(B)*, 제23권 제5호, 1996, pp. 530-539.

[2] 계승균, 표절소고, 인터넷법률신문, 2000년 6월 8일.

[3] 구글 학술정보, http://scholar.google.co.kr.

[4] 나일주, 웹 기반교육, *교육과학사*, 1999.

[5] 류창건, 김형준, 박병준, 최혜정, 조환규, "한글 말뭉치를 이용한 한글 표절 탐색 모델 개발", 2007년도 한국정보과학회 가을 학술발표논문집, 제34권, 제2호, 2007, pp. 58-59.

[6] 박상위, 오정석, 이상호, "메타 검색엔진을 위한 HTML 문서변경탐지기의 설계 및 구현", *정보처리학회논문지D*, 제9권 제3호, 2002.

[7] 박찬정, 임화경, 지은림, "웹을 활용한 수업에서 강의평가 문항분석", 한국정보과학회 학술대회 발표자료집, 2001.

[8] 블로그얌, 2008년도 4월 블로그 이용실적 통계, http://www.blogyam.co.kr/.

[9] 손기락, 문승미, "계층적 군집화 기법을 이용한 소스코드 표절검사", *한국정보교육학회지*, 제11권 제1호, 2007, pp. 91-98.

[10] 오동석, "표절, 침묵의 카르텔과 윤리의 침묵", *인물과 사상*, 제102권, 2006, pp. 65-75.

[11] 임정훈, "인터넷을 활용한 가상수업에서의 교수-학습 활동 및 교육효과 연구", *교육공학연구*, 제14권 제1호, 1998.

[12] 임해창, 최성원, 우연문, 문서의 표절 검사 방법, 특허출원, 2006.

[13] 저작권 위원회, http://www.copyright.or.kr.

[14] 홍윤선, 조선옥, "효과적인 e-러닝 시스템 구축을 위한 과제물 표절 검사", *한국콘텐츠학회/한국통신학회 종합학술대회 논문집*, 제1권 제2호, 2003, pp. 53-59.

[15] 황인수, "e-러닝에서 학습자의 사전동기와 수강관련요인이 강의평가에 미치는 영향에 관한 연구", *정보기술응용연구*, 제15권 제2호, 2008, pp. 33-47.

[16] Beasley, J. D., "The Impact of Technology

on Plagiarism Prevention and Detection", *Plagiarism : Prevention, Practice and Policies Conference*, 2004.

[17] EVE2, http://www.canexus.com/eve.

[18] Glatt Plagiarism Services, http://www.plagiarism. com.

[19] HAM, http://nlp.kookmin.ac.kr/HAM/kor/ham-intr.html.

[20] Heberling, M., "Maintaining Academic Integrity in Online Education", *Online Journal of Distance Learning Administration*, Vol. 5, No. 1, 2002.

[21] Iyer, P., and Singh, A., "Document Similarity Analysis for a Plagiarism Detection Systems", *2nd Indian International Conference on Artificial Intelligence*, 2005, pp. 2534-2544.

[22] Kushmerick, N., Weld, D., and Doorenbos, R., "Wrapper Induction for Information Extraction", *International Joint Conference on Artificial Intelligent*, 1997, pp. 729-735.

[23] Liu, Y. T., Zhang, H. R., Chen, T. W., and Teng, W. G., "Extending Web Search for Online Plagiarism Detection", *1-4244-45004/07 IEEE*, 2007.

[24] Manber, U., "Finding Similar Files in a Aarge File System", *Winter USENIX Technical Conference*, San Francisco, CA, USA., 1994.

[25] McGowan, U., "Plagiarism Detection and Prevention : Are We Putting the Cart Before The Horse?", *Proceedings HERDSA Conference*, 2005.

[26] Mulcahy, S., and Goodacre, C., "Opening Pandora's Box of Academic Integrity : Using Plagiarism Dtection Software", *Proceedings of the 21st ASCILITE Conference*, Perth, 2004, pp. 688-696.

[27] Niezgoda, S., and Way, T. P., "SNITCH : a Software Tool for Detecting Cut and Paste Plagiarism", *SIGCSE Technical Symposium (SIGCSE)*, 2006.

[28] Ottenstein, K. J., "An Algorithmic Approach to The Detection and Prevention of Plagiarism", *CSD-TR 200,* Purdue University, 1976.

[29] Savage, S., "Staff and Student Responses to a Trial of Turnitin Plagiarism Detection Software", *Proceedings of the Australian Universities Quality Forum*, 2004.

[30] Scanlon, P. M. and Neumann, D. R., "Internet Plagiarism Among College Students", *Journal of College Student Development*, Vol. 43, No. 3, 2002, pp. 374-385.

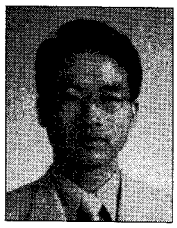
[31] TurnItIn, http://www.turnitin.com.

[32] Vernon, R., Bigna, S., and Smith, M., "Plagiarism and the Web", *Journal of Social Work Education*, Vol. 37, No. 1, 2001, pp. 193-196.

[33] Weir, G. R., Gordon, M. A., and MacGregor, G., "Work in Progress-Technology in Plagiarism Detection and Management", *34th ASEE/IEEE Frontiers in Education Conference*, 2004.

[34] Yeo, S. and Williams, M., "Turnitin : A tool for teachers", *17th Annual Teaching Learning Forum*, 2008.

저자소개



황 인 수

전주대학교 미디어정보학부 정보시스템 전공의 교수로서, 정보통신원장과 교수학습지원센터장을 겸직하고 있다. 고려대학교 경영학과를 졸업하고 동대학원에서 경영정보시스템을 전공하여 석사 및 박사학위를 취득하였으며, 산업연구원(KIET) 물류·유통연구센터의 연구원을 역임하였다. 주요 관심분야는 e-러닝, 인터넷 정보검색 에이전트, 데이터마이닝 등이다.

◈ 이 논문은 2009년 01월 12일 접수하여 1차 수정을 거쳐 2009년 03월 13일 게재확정되었습니다.